# TOWARDS NATURAL HUMAN-AI INTERACTIONS IN VISION AND LANGUAGE

A Dissertation
Presented to
The Academic Faculty

by

Arjun Chandrasekaran

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing

Georgia Institute of Technology
December 2019

# TOWARDS NATURAL HUMAN-AI INTERACTIONS IN VISION AND LANGUAGE

Approved by:

Professor Devi Parikh, Advisor
School of Interactive Computing
*Georgia Institute of Technology*

Professor Dhruv Batra
College of Computing
*Georgia Institute of Technology*

Professor Sonia Chernova
College of Computing
*Georgia Institute of Technology*

Professor Mark Riedl
College of Computing
*Georgia Institute of Technology*

Professor Mohit Bansal
Department of Computer Science
*University of North Carolina at Chapel Hill*

Date Approved: 1 November 2019

*To my family,*

*for their cheer and grit through tumult.*

# ACKNOWLEDGEMENTS

Thanks to all my collaborators and committee members for their interest, expertise, feedback and support that brought the work in this dissertation to fruition.

Several people have contributed to my progress in various ways. I express my profound gratitude to:

Devi Parikh for her exceptional acumen, probing questions and unwavering diligence. And for inviting me into the world of academic research.

Dhruv Batra for his inspiring vision, proactive leadership and commitment to students' success. And for being a role model.

Mohit Bansal for his energy and expertise. And for graciously guiding a fledgling researcher to flight.

Ashwin Kalyan for his singular values and opinions. For being a *nakama*.

Ramakrishna Vedantam, my first researcher friend. For introducing me to the PhD life. And vim!

Ramprasaath Selvaraju for helping me train my first deep neural network. And for the midnight Gatorade runs.

Harsh Agrawal and Abhishek Das for the fun times as roommates. And for the memorable house motto describing criteria for sleep.

Yash Goyal for the reflections, introspections and nourishing conversations.

Aishwarya Agrawal for the antics, banter, dancing, and movie plans that span years.

Prithvijit Chattopadhyay for coffee breaks, contemplations, confessions, and spirited renditions of vintage songs.

Viraj Prabhu for his unassuming charm and being a kindred spirit.

Senthil Purushwalkam, Shrenik Lad, Deshraj Yadav, Samyak Datta, Rishabh Jain,

Ayush Shrivastava, Karan Desai, for the birthday parties, nocturnal conversations and camaraderie.

Clint Solomon, Pooja Harekoppa, Azam Moosavi, Pavan Rangudu, Yash Goyal and Aishwarya Agrawal for many bleary-eyed discussions, and long evenings spent on coursework.

Stefan Lee, for his patience, understanding and insatiable willingness to help.

Jiasen Lu, Jianwei Yang, Michael Cogswell for the fascinating conversations about research, life, and culture. And of course, for Terraforming Mars.

Renee Jamieson for her whole-hearted commitment to students and unconditional support.

Members of the Visual Intelligence lab, Machine Learning & Perception lab, and excellent peers in the ML@GT community for the brainstorming, white-board discussions, feedback and friendship.

Anitha Kannan for her remarkable insight and dedication as a mentor.

Chen Yu for his enthusiasm for research, soccer and sports analogies. For the opportunity and generous guidance in a new field of research.

Sonia Chernova, Bevil Conway and Michael Black for their excellent, candid advice.

Matthew Goodrum for being a partner in volleyball and conversation, which enlivened many evenings.

Blacksburg and its warm people, for making me feel at home in an alien country.

My wonderful roommates through the years.

Srivathsan Iyengar for his selfless advice and backing as a manager and friend.

Sujatha *athai* and Prabhu *athimber* for their support and encouragement.

Kamalahasan *chithapa* for championing me through the years.

My father Narayana Chandrasekaran for his wise and bold guidance, my mother Anuradha Chandrasekaran for her love, my sister Amala Chandrasekaran and brother-in-law Bharathwaj Nandakumar for their solidarity.

God, for everything.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# Thesis statement

Neural networks and humans are successful at modeling aspects of each other – neural networks can model humor and narrative, humans can predict responses of neural networks and collaborate with them more effectively when their decisions are explained. Specifically,

- neural network models inspired from cognitive theories are competitive with appropriately constrained lay persons on visual and contextual humor tasks
- a neural network model previously trained on stories, outperforms baselines in a cognitive task involving narrative in real-life events
- lay-persons learn to predict outputs and failures of deep networks better after familiarization with the model
- lay-persons utilize interpretable visual and text explanations about a network's response to improve collaborative performance.

# SUMMARY

Inter-human interaction is a rich form of communication. Human interactions typically leverage a good theory of mind, involve pragmatics, story-telling, humor, sarcasm, empathy, sympathy, etc. Recently, we have seen a tremendous increase in the frequency and the modalities through which humans interact with AI. Despite this, current human-AI interactions lack many of these features that characterize inter-human interactions. Towards the goal of developing AI that can interact with humans naturally (similar to other humans), I take a two-pronged approach that involves investigating the ways in which both the AI and the human can adapt to each other's characteristics and capabilities. In my research, I study aspects of human interactions, such as humor, story-telling, and the humans' abilities to understand and collaborate with an AI. Specifically, in the vision and language modalities,

1. In an effort to improve the AI's capabilities to adapt its interactions to a human, we build computational models for (i) humor manifested in static images, (ii) contextual, multi-modal humor, and (iii) temporal understanding of the elements of a story.

2. In an effort to improve the capabilities of a collaborative human-AI team, we study (i) a lay person's predictions regarding the behavior of an AI in a situation, (ii) the extent to which interpretable explanations from an AI can improve performance of a human-AI team.

Through this work, I demonstrate that aspects of human interactions (such as certain forms of humor and story-telling) can be modeled with reasonable success using computational models that utilize neural networks. On the other hand, I also show

that a lay person can successfully predict the outputs and failures of a deep neural network. Finally, I present evidence that suggests that a lay person who has access to interpretable explanations from the model, can collaborate more effectively with a neural network on a goal-driven task.

# CHAPTER I

# INTRODUCTION

The recent surge of advances in deep learning, has resulted in the large-scale use of AI approaches across a diverse set of applications. A large fraction of these applications involve interfacing with humans, e.g., in conversational bots, web search, etc. While progress in research is increasing the number of applications where AI is used, it is also important to improve the usability of these AI systems. My research goal is to develop AI systems that can interact in a manner that is natural to humans. In an ideal scenario, humans should find interacting with the AI to be at least as easy as interacting with other humans (such as in Fig. 1).

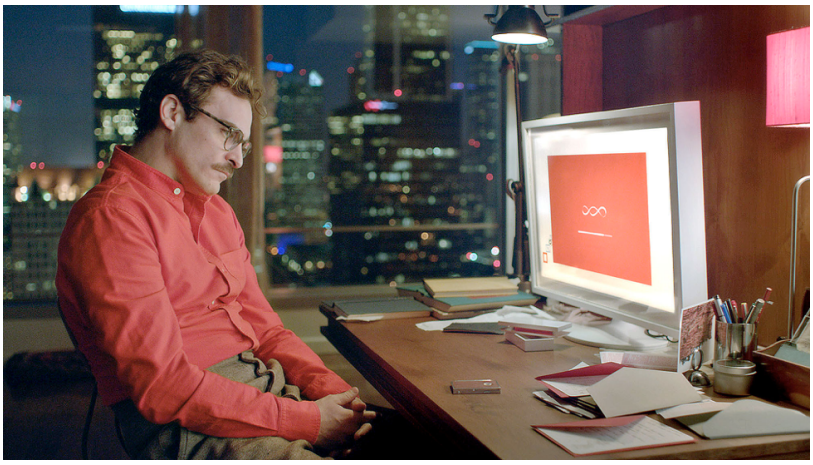
Towards the end of developing AI that interacts naturally with humans, we first observe natural interactions among humans. Inter-human interaction is rich – we typically have a good theory of mind, our interactions involve pragmatics, humor, sarcasm, empathy, sympathy, our communication often involves story telling, etc. Current human-AI interactions lack many of these features that characterize inter-human interactions. In this dissertation, I take steps towards studying aspects of humor, story-telling, and theory of (AI's) mind. Specifically,

1. We build computational models for humor manifested in static images (described in Sec. 1.1), and contextual, multi-modal humor (described in Sec. 1.2).

2. We introduce a picture-sequencing task where a computational model learns the correct temporal order of events in a story (Sec. 1.3).

3. We evaluate the predictability of a deep neural network to a lay person (described in Sec. 1.4).

(a) Tony Stark issuing instructions in natural language to his AI engineering assistant Jarvis, in *Iron Man*. Source: `https://digitaltimes.com.mm/tag/artificial-intelligence/`



(b) Theodore Twombly falls in love with Samantha (an AI) over a number of conversations, in *Her*. Source: `www.readthespirit.com/visual-parables`

**Figure 1:** Fictional examples of AI that people can interact with naturally. These AI have a sense of humor, can understand and form narrative, and their behavior is reasonably predictable by the people interacting with them.

4. We evaluate the role of an interpretable explanation from the model in improving performance of a human-AI team (described in Sec. 1.5).

In all of these works, we focus on interactions that are in the two primary modes of communication in humans – vision and language. In the following sections, we describe each of these works in further detail.

## *1.1 Visual humor*

Humor is an essential component of inter-human interactions. Indeed, research has found that humor is critical in social interactions even amongst monkeys. Humor is prevalent in our everyday lives. It manifests itself in different forms (see Fig. 2) that each includes a subset of modalities – slapstick humor by clowns, purely visual cartoons on television, cartoons with text in newspapers, television series involving dialogue, stand-up comedy shows, and textual humor in literature. In my first work, we focus on understanding and predicting purely visual humor, i.e., humor that is manifested in static scenes.

In Chapter 3 on visual humor, we present two tasks that we posit, demonstrates an algorithm's understanding of aspects of visual humor at differing levels of detail. First, given a scene, the algorithm predicts the extent to which it is funny. This involves an understanding of humor at the scene-level. Second, given a scene, the algorithm changes aspects of the scene to make the scene more/less funny. This task requires an understanding of humor at a more fine-grained level, i.e., the algorithm must be able to identify specific elements in the scene that make it funny. We then present computational models of visual humor that solve the above two tasks.

Understanding and predicting visual humor in static scenes involves a number of challenges such as detecting objects in potentially novel contexts and other low-level computer vision tasks. We side-step these challenges by leveraging abstract scenes [257, 255], which are trivially densely annotated via the interface used to create


**Figure 2:** Humor manifests itself in multiple forms in our everyday lives, spanning sources and modalities.

the scenes. Fig. 3 shows an example scene with trivial annotations. Working with abstract scenes allows us to directly focus and address the challenge of understanding the higher level semantics of the scene without having to engage with the challenges of solving low-level computer vision.

Leveraging abstract scenes, we crowd-source funny scenes and analyze the specific techniques that people use to make the scenes funny. We also ask people to identify the set of objects in a scene that contribute to humor and replace these with other objects so that the resulting scene is not funny anymore. We use these two datasets as training data and train computational models for visual humor.

In human studies, we find that our models of visual humor that alter the funniness of a scene perform well. The model identifies objects that contribute to humor in a scene with reasonable success. Further, the model replaces these objects with other 'boring' objects, effectively eliminating the humor in a given funny scene about 95% of the time. Our model that introduces humor by altering a boring scene is compared

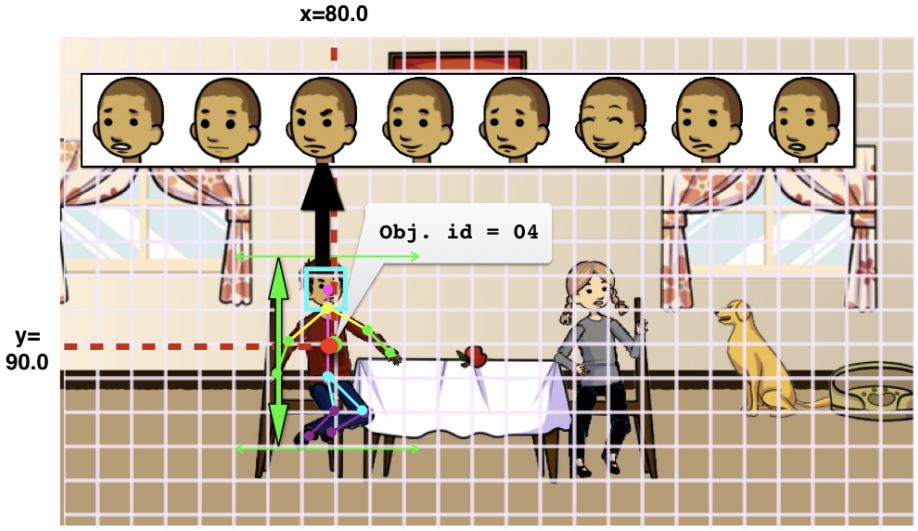



**Figure 3:** Abstract scenes are trivially annotated with the identities of objects, their positions, pose, depth, expressions of humans, etc.

with the originally funny scene created by the human in a setup similar to a Turing test. People find the scenes from our model to be funnier than the human scenes about 28% of the time.

## *1.2 Multi-modal humor*

In social interactions, humans are often witty in context. I.e., given a novel situation or scenario, a person often makes a remark that is witty in the given context. Such contextual humor forms the majority of humor in professional contexts, and a large part of humor in social contexts that do not involve recounting 'canned' (memorized) jokes. An extreme notion of this form of humor is characterized by 'improv' performances by stand-up comedians. This is challenging even for most people as it involves making an association in the given novel context that other people find unexpected and witty.

The ability to be contextually witty can be considered one of the most challenging problems in humor. Unlike applications that involve generating 'canned' humor, algorithms that are contextually witty not only need to understand what constitutes

a witty remark in isolation, but also accurately understand the current context, and the role of the new remark given the current context. Apart from being a worthy ideal in the pursuit of strong AI, contextual humor is also an extremely valuable skill for practically applicable AI systems such as conversational agents, AI assistants, and numerous other applications. As a concrete example, an algorithm could suggest a witty response to a friend's picture posted on social media.

In Chapter 4, we attempt to tackle this challenging problem by considering an extremely simple form of humor created using puns. Puns, which are words that sound the same but mean different things, intrinsically have the tendency to introduce ambiguity in the meaning of a sentence. Inspired by the incongruity-resolution theory of humor [215], we attempt to leverage the alternate meaning of a sentence to introduce an incongruity in the perceiver's mind. On closer inspection and further thought, the perceiver (ideally) *resolves* the alternate meaning of the sentence and finds the utterance witty. Fig. 4 presents a concrete example.

We develop two types of computational models that can describe a given possibly 'boring' image in a witty manner, by utilizing puns. Via human studies, we find that people find descriptions written by humans for an image to be wittier compared to witty descriptions from our model, in almost all instances. While this is unsurprising, it is interesting to note that when the humans are constrained to use the same pun words and linguistic style as the model's descriptions, people find the descriptions from the model to be slightly wittier than the constrained humans' descriptions. This identifies areas of research such as generating natural language descriptions that mimic the range of variations and creativity of human descriptions such as the ability to effectively utilize a vast vocabulary that includes rare words, the ability to generate descriptions of varying styles, and lengths.

**Figure 4:** ***Description****: The wedding was so emotional. Even the cake was in tiers.* The image which depicts the emotional women who are in tears, might initially bias a perceiver towards interpreting the description to mean that the cake was in tears. This creates an incongruity in the perceiver's mind. On further observation and thought, it becomes clear that the alternate interpretation of the wedding cake being in tiers, makes more sense. This leads to resolution, and an appreciation of wit.

**Figure 5:** An example picture sequencing task where a student attempts to sort the given jumbled set of pictures, in the temporally correct order of events.

## *1.3 Narrative*

Picture sequencing is a popular task that children in kindergarten are asked to perform as an evaluation of cognitive abilities. Tasks such as that presented in Fig. 5 aim to evaluate children's perception of the progression of events in time, the sequence of steps required to perform a task, and beliefs held by characters in the story. An accurate understanding of the progression of events, for instance, is necessary for interactions with other people who assume that the other person has a 'temporal common-sense'. For instance, when a person mentions that they went to school, it implies that they woke up and likely that they brushed their teeth, changed, etc. It is important for AI systems that interact with humans, to be able to understand such implicit facts that are not explicitly stated by people.

Since ancient times, humans often communicate in the form of stories or narrative. A friend might recount an event that occurred the previous day in the form of a story, and a scientist might attempt to present their research with a clear narrative. Given its ubiquitous presence in human expression, it is essential that AI systems also have the ability to understand aspects of narrative. In addition, an understanding of narrative will likely be useful in practical applications. For instance, an algorithm that understands the temporal ordering of events can automatically create a narrative from disparate individual events of a story.

In the context of narrative, in Chapter 5, we propose the following concrete task – given a jumbled series of events belonging to a story, sort them into the temporally correct sequence. We consider stories that are comprised of static images, each associated with a story-like caption. We train machine learning models to sort the given jumbled stories, and find that even relatively simple models can learn the three-part structure of stories. We observe that models tend to exploit the linguistic markers identifying the different parts of the story, e.g., story beginnings typically are described using words such as, 'reunion', 'carnival', 'winterskate', etc. which describe

the overall event, and story endings using 'overall', 'lastly', 'returned', etc. We find that models that are more discriminative, i.e., those that compare pairs of elements of the story perform the task of sequencing more accurately.

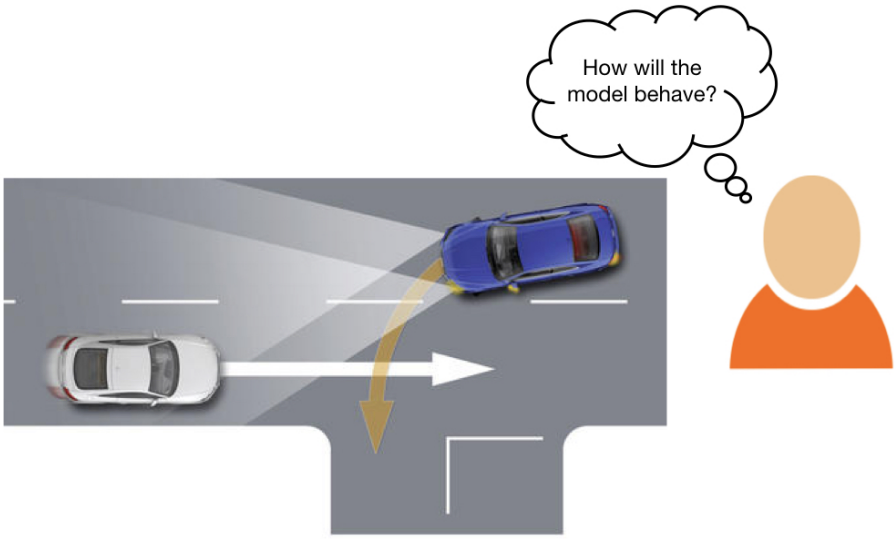
## *1.4 Predictability*

We routinely use technology to achieve a wide variety of tasks. An important consideration while using tools is for the behavior of these tools to be predictable. This notion has been formalized in engineering disciplines which discuss the standard operating conditions and response characteristics of mechanical and electronic elements of a system. The predictability of such systems is of paramount importance to prevent suboptimal outcomes. With the recent paradigm shift in machine learning and its application to a wide range of domains, it is important that we focus on predictability of deep learning models.

The predictability of the systems that we interact and collaborate with is not just restricted to technology – it also forms an important part of social interaction with other people. Theory of Mind (ToM) is the ability to attribute mental states (beliefs, intents, knowledge, perspectives, etc.) to others and recognize that these mental states may differ from one's own. ToM is critical to effective communication and to teams demonstrating higher collective performance. To effectively leverage the progress in Artificial Intelligence (AI) to make our lives more productive, it is important for humans and AI to work well together in a team.

Consider the two safety-critical applications of AI systems in Fig. 6. It is extremely important that the human collaborators in these human-AI teams can adequately predict the behavior of the AI systems. To this end, in our work on predictability, we evaluate the extent to which a lay person can predict the behavior of a deep neural network in a given context. Specifically, we evaluate the predictability of a visual question answering (VQA) [131] model.


(a) A doctor working with a medical diagnosis system is assessing the output from the model. It is important for the doctor to have appropriate trust in such a system to effectively utilize its recommendations in the final diagnosis.




(b) A driver using a driver-assistance system needs to be able to predict the behavior of the model in a given scenario in order to potentially adjust, correct the driving trajectory or adapt to it.

**Figure 6:** Applications where a human is collaborating with an AI to perform a safety-critical task. It is essential that the model is predictable – i.e., it's behavior in a given context can be anticipated by the human.

We first evaluate the extent to which a person can predict if the model will succeed or fail on a given instance. This is important to gauge the appropriate trust of the person in the model. We also study the extent to which a person can predict the exact response of the model. This is a more fine-grained evaluation of the predictability of the model. We find that the model's behavior is more predictable to a lay person than random chance, and that the model's predictability improves as the person is familiarized with the model before-hand, by allowing the human to observe the model's responses to a fixed set of input examples.

While deep neural networks are popular due to their superior performance on a number of tasks, they are infamous for being uninterpretable. To address this issue, there are a number of works that aim to improve the interpretability of deep networks [250, 85, 191, 92, 203, 253, 11, 246]. In this work, we also evaluate the extent to which a few of these approaches towards interpretability influence the predictability of the VQA model. Interestingly, we find that the explanations to the human (in an effort to make the model more interpretable) do not in fact make the model more predictable, when the human is already familiar with the model.

## *1.5 Explanations in human-AI teams*

We evaluated the predictability of a VQA model via two proxy tasks in our work described in Sec. 1.4. While this provides a useful signal regarding the predictability of the model and the utility of explanation modalities in isolation, it is unclear to what extent a person might be able to leverage the predictability of a model in a downstream task. We propose to evaluate this by formulating a co-operative task where the human is required to collaborate with the model in order to achieve a goal.

We leverage the game of GuessWhich, introduced in our previous work [41] as a testbed to evaluate the human-AI team in a goal-driven co-operative task. In GuessWhich the human attempts to guess the 'secret' image from a pool of $N$ images

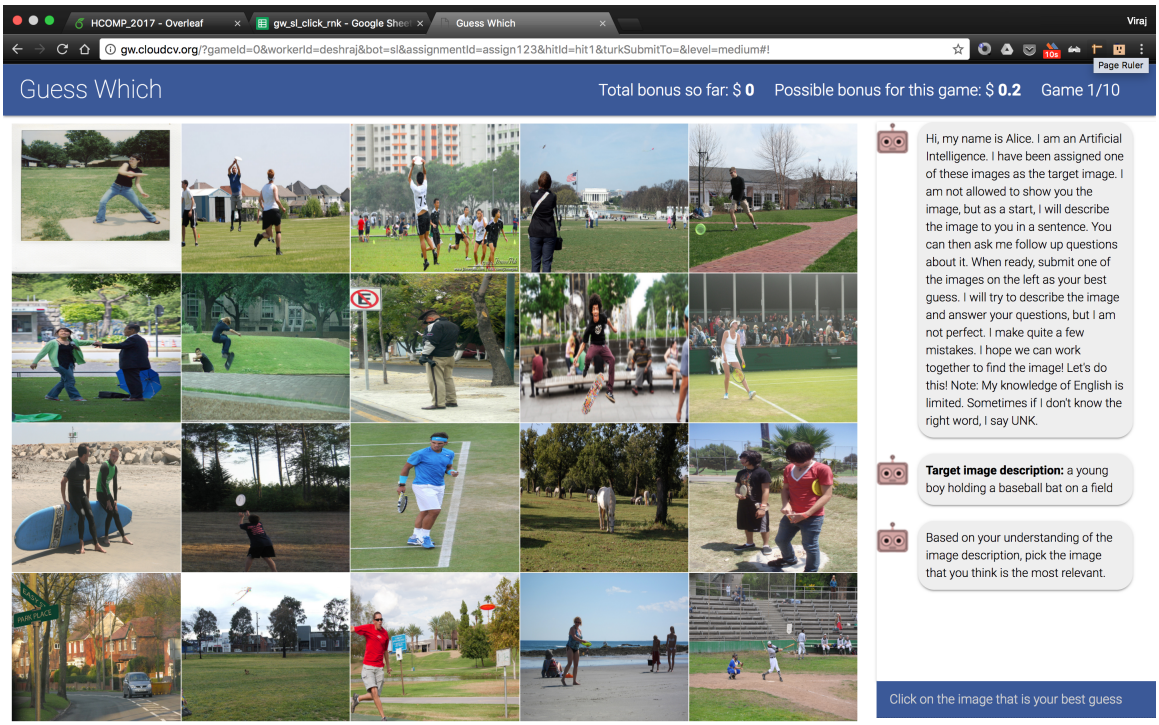



**Figure 7:** Screenshot of the GuessWhich interface.

(similar to the game of '20 questions'). The human subject asks the AI (a VQA model such as [131] or visual dialog model such as [54]) a question about the secret image, and receives a response from the model. The subject now has the opportunity to ask a subsequent question, after taking into account the model's response to their previous question. After $R$ rounds of question-answering, the human subject makes a final guess of the secret image. The performance of the human-AI team is measured by the number of guesses it takes the human to correctly identify the secret image after a fixed number of dialog rounds with the AI. A screenshot of the GuessWhich interface used in [41] is shown in Fig. 7.

In this work, we evaluate the role of explanations in improving predictability of the model in the context of the downstream task of GuessWhich. We add features to the GuessWhich interface to also provide the subject explanations regarding the answer, in addition to the answer to the answer itself. Similar to GuessWhich, given a pool of images, the subject begins by asking a question about the secret image.

Following the answer from a VQA model, the subject attempts to guess the secret image. After making the guess without explanations, the subject is then provided with an explanation regarding the model's answer to the user's question. The subject attempts to guess the secret image again, with the opportunity to take into account the explanation from the model. We hypothesize that an improvement in accuracy of the human's guesses, demonstrates the utility of the explanation in increasing the predictability of the model to the human subject.

We perform experiments with Grad-CAM [203], a saliency based visual explanation, and provide preliminary results with text-explanations. We find that while Grad-CAM does not to improve the predictability of the model in this goal-driven task, text-explanations appear to hold promise. We describe this work in further detail in Chapter 7.

## *1.6 Contributions*

In this dissertation, we describe our work which attempts to make progress towards the goal of achieving natural human-AI interactions. To that end, we develop and implement computational models of visual humor, contextual multi-modal humor, and narrative. In addition, via human studies, we evaluate the extent to which an AI is predictable to a human teammate. We also propose to evaluate the role of explanation modalities in a concrete, goal-driven, co-operative human-AI task. Our research contributions in each of these works, are described below:

1. To the best of my knowledge, our work on visual humor is the first work that deals with understanding and building computational models for visual humor (described in Chapter 3). We collect two abstract scene datasets consisting of scenes created by humans which are publicly available. Further, we analyze the different sources of humor techniques depicted in the AVH dataset via human studies. We propose two tasks that evaluate an understanding of visual humor,

and implement computational models that perform well on the tasks. In human evaluation, people find the scenes made funny by our approach to be funnier than the originally funny human scene about 28% of the time.

2. To the best of my knowledge, our work on contextual humor is the first work that tackles the challenging problem of producing a witty natural language remark in an everyday (boring) context. In our work, we present two novel models to produce witty (pun-based) captions for a novel (likely boring) image. In human evaluations, we find that our approach generates descriptions that are just as witty as humans who are constrained to use similar pun words and style as our approach.

3. We introduce the task of visual story sequencing as a concrete way to evaluate an algorithm's understanding of the temporal order of events in a story. We implement two approaches to solve the task that utilize both text and image features, and find that our computational models perform well compared to other relevant approaches.

4. We introduce two concrete proxy tasks to evaluate the predictability of a deep neural network. We implement an interface that allows human subjects on a crowd-sourcing platform to get familiarized with the deep model, and find that the predictability of the model improves. We also evaluate a few existing explanation modalities and surprisingly, find that they do not make the model more predictable to a lay person. Thus, predictability can serve as a concrete, goal-driven measure of the utility of explanations to a lay person.

5. We evaluate the role of explanations in a concrete goal-driven co-operative task involving a human collaborating with a deep neural network. This work aims to measure the utility of interpretability approaches in a goal-driven setting that involves a live interaction of a human with an AI.

We next discuss the research related to the work presented in this dissertation, followed by a detailed discussion of each of the research topics that were described in brief, above. We then conclude the dissertation and provide appendices that contain further details.

# CHAPTER II

# RELATED WORK

## *2.1 Visual humor*

**Humor Theories.** Humor has been a topic of study since the time of Plato [175], Aristotle [7] and Bharata [18]. Over the years, philosophical studies and psychological research have sought to explain why we laugh. There are three theories of humor [238] that are popular in contemporary academic literature. According to the incongruity theory, a perceiver encounters an incongruity when expectations about the stimulus are violated [134]. The two stage model of humor [215] further states that the process of discarding prior assumptions and reinterpreting the incongruity in a new context (resolution) is crucial to the comprehension of humor. Superiority theory suggests that the misfortunes of others which reflects our own superiority is a source of humor [161]. According to the relief theory, humor is the release of pent-up tension or mental energy. Feelings of hostility, aggression, or sexuality that are expressed bypassing any societal norms are said to be enjoyed [74].

Previous attempts to characterize the stimuli that induce humor have mostly dealt with linguistic or verbal humor [145] *e.g.*, script-based semantic theory of humor [182] and its revised version, the general theory of verbal humor [197].

**Computational Models of Humor.** A number of computational models are developed to recognize language-based humor e.g., one-liners [147], sarcasm [55] and *knock-knock* jokes [219]. Other work in this area includes exploring features of humorous texts that help detection of humor [146], and identifying the set of words or phrases in a sentence that could contribute to humor [247].

Some computational humor models that generate verbal humor are JAPE [23]

which is a pun-based riddle generating program, HAHAcronym [214] which is an automatic funny acronym generator, and an unsupervised model that produces "*I like my X like I like my Y, Z*" jokes [173]. While the above works investigate detection and generation of verbal humor, in this work we deal purely with *visual* humor.

Recent works predict the best text to go along with a given (presumably funny) raw image such as a meme [231] or a cartoon [204]. In addition, Radev et al. [180] develop unsupervised methods to rank funniness of captions for a cartoon. They also analyze the characteristics of the funniest captions. Unlike our work, these works do not predict whether a *scene* is funny or which components of the scene contribute to the humor.

Buijzen and Valkenburg [33] analyze humorous commercials to develop and investigate a typology of humor. Our contributions are different as we study the sources of humor in static images, as opposed to audiovisual media. To the best of my knowledge, ours is the first work to study *visual* humor in a computational framework.

**Human Perception of Images.** A number of works investigate the intrinsic characteristics of an image that influence human perception e.g., memorability [101], popularity [112], visual interestingness [88], and virality [58]. While there may exist a correlation between funniness and the above properties (e.g., funny images may tend to be memorable or popular), they are still distinct (all memorable or popular images may not be funny). In this work, we study what content in a scene causes people to perceive it as funny, and explore a method of altering the funniness of a scene.

**Learning from Visual Abstraction.** Visual abstractions have been used to explore high-level semantic scene understanding tasks like identifying visual features that are semantically important [257, 255], learning mappings between visual features and text [258], learning visually grounded word embeddings [119], modeling fine-grained interactions between pairs of people [6], and learning (temporal and static) common sense [73, 126, 223]. In this work, we use abstract scenes to understand the semantics

in a scene that cause humor, a problem that has not been studied before.

## *2.2 Multi-modal humor*

**Humor theory.** General Theory of Verbal Humor [9] characterizes linguistic stimuli that induce humor but implementing computational models of it requires severely restricting its assumptions [21].

**Puns.** Zwicky [260] classify puns as *perfect* (pronounced exactly the same) or *imperfect* (pronounced differently). Similarly, Pepicello [172] categorize riddles based on the linguistic ambiguity that they exploit – phonological, morphological or syntactic. Kao et al. [108] formalize the notion of incongruity in puns and use a probabilistic model to evaluate the funniness of a sentence. Jaech et al. [102] learn phone-edit distances to predict the counterpart, given a pun by drawing from automatic speech recognition techniques. In contrast, we augment a web-scraped list of puns using an existing model of pronunciation similarity.

**Generating textual humor.** JAPE [23] also uses phonological ambiguity to generate pun-based riddles. While our task involves producing free-form responses to a novel stimulus, JAPE produces stand-alone "canned" jokes. HAHAcronym [214] generates a funny expansion of a given acronym. Unlike our work, HAHAcronym operates on text, and is limited to producing sets of words. [173] develop an unsupervised model that produces jokes of the form, "*I like my X like I like my Y, Z*" .

**Generating multi-modal humor.** Wang and Wen [231] predict a meme's text based on a given funny image. Similarly, Shahaf et al. [205] and Radev et al. [180] learn to rank cartoon captions based on their funniness. Unlike typical, boring images in our task, memes and cartoons are images that are already funny or atypical. E.g., "LOL-cats" (funny cat photos), "Bieber-memes" (modified pictures of Justin Bieber), cartoons with talking animals, etc. Chandrasekaran [38] alter an abstract scene to

make it more funny. In comparison, our task is to generate witty natural language remarks for a novel image.

**Poetry generation.** Although our tasks are different, our generation approach is conceptually similar to Ghazvininejad et al. [80] who produce poetry, given a topic. While they also generate and score a set of candidates, their approach involves many more constraints and utilizes a finite state acceptor unlike our approach which enforces constraints during beam search of the RNN decoder.

## *2.3 Narrative*

Temporal ordering has a rich history in NLP research. Scripts [202], and more recently, narrative chains [35] contain information about the participants and causal relationships between events that enable the understanding of stories. A number of works [136, 138, 26] learn temporal relations and properties of news events from the dense, expert-annotated TimeBank corpus [179]. In our work, however, we use multi-modal story data that has no temporal annotations.

A number of works also reason about temporal ordering by using manually defined linguistic cues [235, 170, 123, 93, 111]. Our approach uses neural networks to avoid feature design for learning temporal ordering.

Recent works [158, 157] learn distributed representations for predicates in a sentence for the tasks of event ordering and cloze evaluation. Unlike their work, our approach makes use of multi-modal data with *free-form* natural language text to learn event embeddings. Further, our models are trained end-to-end while their pipelined approach involves parsing and extracting verb frames from each sentence, where errors may propagate from one module to the next (as discussed in Section 5.2.3).

Chen et al., [44] use a generalized Mallows model for modeling sequences for coherence within single documents. Their approach may also be applicable to our task. Recently, Mostafazadeh [160] presented the "ROCStories" dataset of 5-sentence

stories with stereotypical causal and temporal relations between events. In our work though, we make use of a multi-modal story-dataset that contains *both* images and associated story-like captions.

Some works in vision [174, 15] also temporally order images; typically by finding correspondences between multiple images of the same scene using geometry-based approaches. Similarly, Choi et al. [48] compose a story out of multiple short video clips. They define metrics based on scene dynamics and coherence, and use dense optical flow and patch-matching. In contrast, our work deals with stories containing potentially visually dissimilar but *semantically* coherent set of images and captions.

A few other recent works [113, 114, 115, 207, 28, 230] summarize hundreds of individual streams of information (images, text, videos) from the web that deal with a single concept or event, to learn a common theme or *storyline* or for *timeline summarization*. Our task, however, is to predict the correct sorting of a given story, which is different from summarization or retrieval. Ramanathan et al. [181] attempt to learn temporal embeddings of video frames in complex events. While their motivation is similar to ours, they deal with sampled frames from a video while we attempt to learn temporal common sense from *multi-modal* stories consisting of a sequence of aligned image-caption pairs.

## *2.4 Predictability*

**Explanations in deep neural networks.** Several works generate explanations based on internal states of a decision process [250, 85], while others generate justifications that are consistent with model outputs [191, 92]. Another popular form of providing explanations is to visualize regions in the input that contribute to a decision – either by explicitly attending to relevant input regions [11, 246], or exposing implicit attention for predictions [203, 253]. In our work, we evaluate three kinds of explanations – introspective, implicit attention, and explicit attention.

**Evaluating explanations.** Several works evaluate the role of explanations in developing trust with users [51, 191] or helping them achieve an end goal [165, 122]. Further, [177] evaluate the interpretability of a simple linear regression based on users' ability to predict failure of the model, similar to our FP task. Our work, however, investigates the role of machine-generated explanations in improving the *predictability* of a VQA model. We also consider an inherently interactive task – VQA, where a person actively queries the AI about an image, which naturally leads to collaborative settings, i.e., human-AI teams. Unlike all the above works, we explore the role of explanations in complex and inherently uninterpretable deep neural networks.

**Failure prediction.** While Bansal et al. [12] and Zhang et al. [251] predict failures of a model using simpler statistical models, we explicitly train a person to do this. We also evaluate the role of familiarity with and without explanation modalities. In addition to predicting the success or failure of an AI agent, we also train humans to more accurately predict the 'behavior', i.e., the actual output of an AI agent.

**Legibility.** Dragan et al. [60] describe the intent-expressiveness of a robot as its trajectory being expressive of its goal. Analogously, we evaluate if explanations of the intermediate states of a VQA model are expressive of its output. Our experiments to measure the predictability is in line with evaluating the legibility of VQA models.

**Humans adapting to technology.** Wang et al. [229] and Pelikan et al. [171] observe humans' strategies while adapting to the limited capabilities of an AI in interactive language games. For instance, in a human-AI game of charades, humans modify strategies such as word selection, turn length, and prosody, humans modify strategies to adapt to the robot's limited perceptive abilities. While these works observe that humans dynamically adapt their behavior to adapt to the agent's limited perceptive abilities, In our work we explicitly measure to what extent humans can form an accurate model of an AI, and the role of familiarization and explanations.

# CHAPTER III

# VISUAL HUMOR

***TARS:*** *[as Cooper repairs him] Settings. General settings. Security settings.*

***Cooper:*** *Honesty, new setting: 95%.*

***TARS:*** *Confirmed. Additional settings.*

***Cooper:*** *Humor, 75%.*

***TARS:*** *Confirmed. Self-destruct sequence in T minus 10, 9...*

***Cooper:*** *Let's make that 60%.*

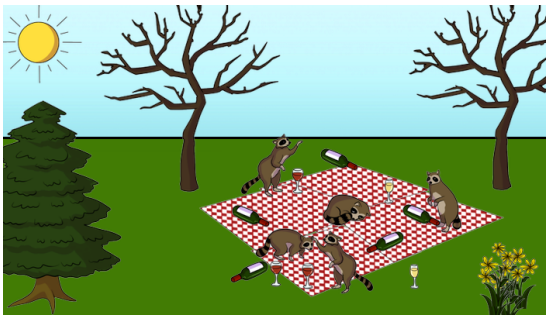
***TARS:*** *Sixty percent, confirmed. Knock, knock.*

***Cooper:*** *You want 55?*

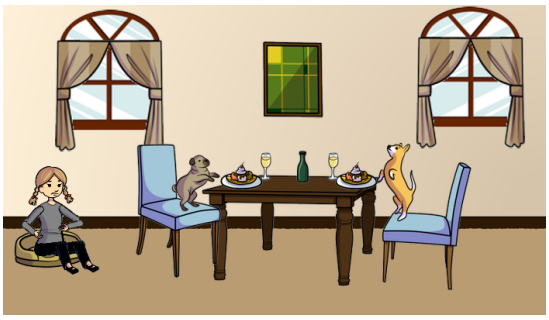
— *Interstellar (film), 2014.*

An adult laughs 18 times a day [139] on average. A good sense of humor is related to communication competence [62, 63], helps raise an individual's social status [199], popularity [81, 140], and helps attract compatible mates [31, 34, 163]. Humor in the workplace improves camaraderie and helps workers cope with daily stresses [176] and loneliness [232]. *fMRI* [193] studies of the brain reveal that humor activates the components of the brain that are involved in reward processing [233]. This probably explains why we actively seek to experience and create humor [156].

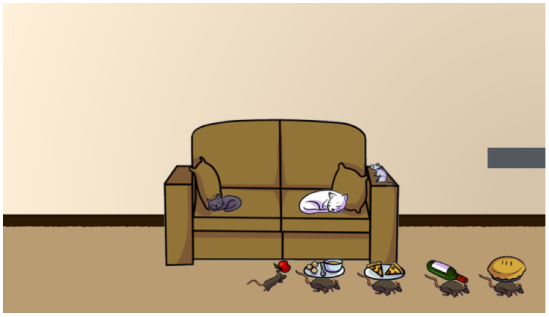
Despite the tremendous impact that humor has on our lives, the lack of a rigorous definition of humor has hindered humor-related research in the past [8, 211]. While verbal humor is better understood today [197, 182], visual humor remains unexplored. As vision and AI researchers we are interested in the following question – what content



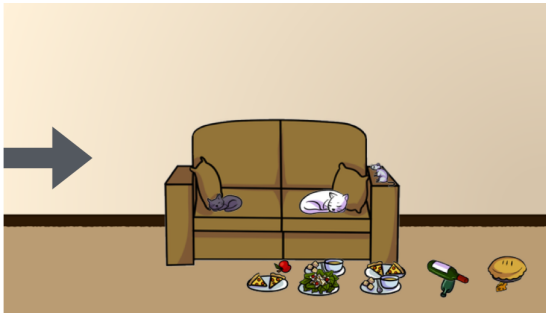
(a) *Funny scene:* Raccoons are drunk at a picnic.



(b) *Funny scene:* Dogs feast while the girl sits in a pet bed.



(c) *Funny scene:* Rats steal food while the cats are asleep.



(d) *Funny Object Replaced (unfunny) counterpart:* Rats in *(c)* are replaced by food.

**Figure 8:** (a), (b) are selected funny scenes in the Abstract Visual Humor dataset. (c) is an originally funny scene in the Funny Object Replaced dataset. The objects contributing to humor in (c) are replaced by a human with other objects, to create an unfunny counterpart.

in an image causes it to be funny? Our work takes a step in the direction of building computational models for visual humor.

Computational visual humor is useful for a number of applications: to create better photo editing tools, smart cameras that pick the right moment to take a (funny) picture, recommendation tools that rate funny pictures higher (say, to post on social media), video summarization tools that summarize only the funny frames, automatically generating funny scenes for entertainment, identifying and catering to personalized humor, *etc*.

As AI systems interact more with humans, it is vital that they understand subtleties of human emotions and expressions. In that sense, being able to identify humor can contribute to their *common sense*.

Understanding visual humor is fraught with challenges such as having to detect all objects in the scene, observing the interactions between objects, and understanding context, which are currently unsolved problems.

In this work, we argue that, by using scenes made from clipart [5, 6, 73, 119, 126, 223, 257, 258], we can study visual humor without having to wait for these detailed recognition problems to be solved. Abstract scenes are inherently densely annotated (*e.g.* all objects and their locations are known), and so enable us to learn fine-grained semantics of a scene that causes it to be funny. In this paper, we collect two datasets of abstract scenes that facilitate the study of humor at both the scene-level (Fig. 8(a), Fig. 8(b)) and the object-level (Fig. 8(c), Fig. 8(d)). We propose a model that predicts how funny a scene is using semantic visual features of the scene such as occurrence of objects, and their relative locations. We also build computational models for a particular source of humor, *i.e.*, humor due to the presence of objects in an unusual context. This source of humor is explained by the *incongruity theory* of humor which states that a playful violation of the subjective expectations of a perceiver causes humor [145]. *E.g.*, Fig. 8(b) is funny because our expectation is that people eat at tables and dogs sit in pet beds and this is violated when we see the roles of people and dogs swapped.

The scene-level Abstract Visual Humor (AVH) dataset contains funny scenes (Fig. 8(a), Fig. 8(b)) and unfunny scenes with human ratings for funniness of each scene. Using the ground truth rating, we demonstrate that we can reliably predict a *funniness score* for a given scene. The object-level Funny Object Replaced (FOR) dataset contains scenes that are originally funny (Fig. 8(c)) and their unfunny counterparts (Fig. 8(d)). The unfunny counterparts are created by humans by replacing objects that contribute to humor such that the scene is not funny anymore. The ground truth of replaced objects is used to train models to alter the funniness of a scene – to make a funny scene unfunny and vice versa. Our models outperform

natural baselines and ablated versions of our system in quantitative evaluation. They also demonstrate good qualitative performance via human studies.

Our main contributions are as follows:

1. We collect two abstract scene datasets consisting of scenes created by humans which are publicly available.

    i. The scene-level Abstract Visual Humor (AVH) dataset consists of funny and unfunny abstract scenes (Section 3.1.2). Each scene also contains a brief explanation of the humor in the scene.

    i. The object-level Funny Object Replaced (FOR) dataset consists of funny scenes and their corresponding unfunny counterparts resulting from object replacement (Section 3.1.3).

1. We analyze the different sources of humor techniques depicted in the AVH dataset via human studies (Section 3.1.2).

1. We learn distributed representations for each object category which encode the context in which an object naturally appears, *i.e.*, in an unfunny setting. (Section 3.2.1).

1. We model two tasks to demonstrate an understanding of visual humor:

    i. Predicting how funny a given scene is (Section 3.3.1).

    i. Automatically altering the funniness of a given scene (Section 3.3.2).

To the best of our knowledge, this is the first work that deals with understanding and building computational models for visual humor.

## *3.1 Datasets*

We introduce two new abstract scenes datasets – the Abstract Visual Humor (AVH) dataset (Section 3.1.2) and the Funny Object Replaced (FOR) dataset (Section 3.1.3)

using the interfaces described in Section 3.1.1. The AVH dataset (Section 3.1.2) consists of both funny and unfunny scenes along with funniness ratings. The FOR dataset (Section 3.1.3) consists of funny scenes and their altered unfunny counterparts. Both the datasets are made publicly available on the project webpage.

### 3.1.1 Abstract Scenes Interface

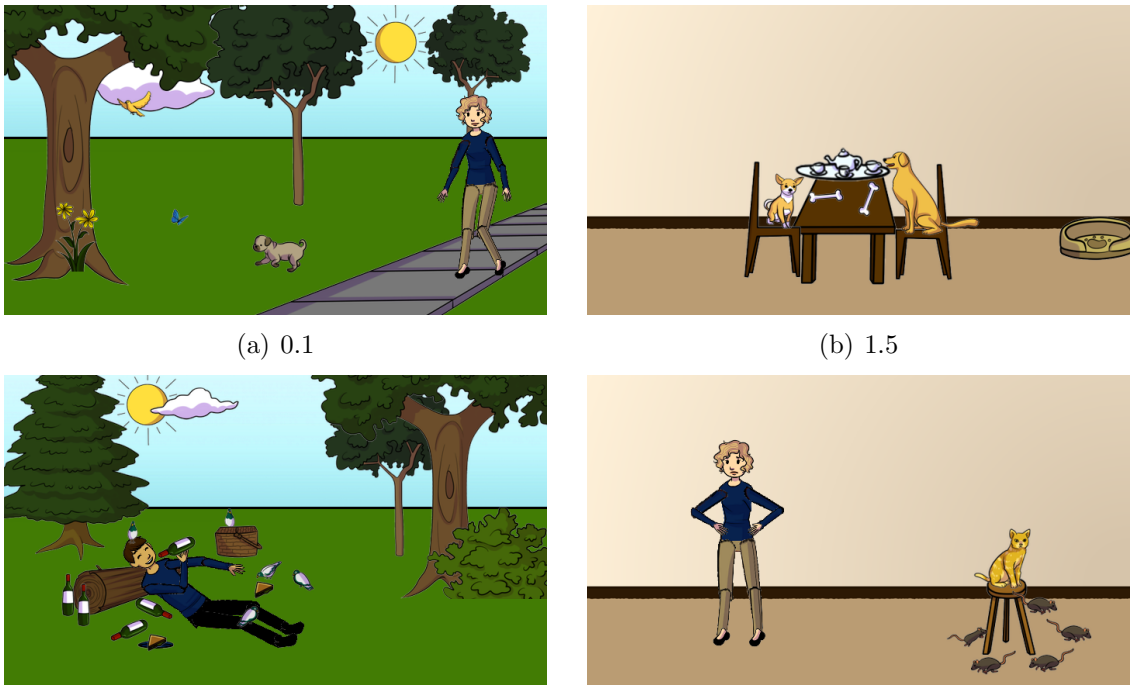
Abstract scenes enable researchers to explore high-level semantics of a scene without waiting for low-level recognition tasks to be solved. We use the clipart interface[1] developed by Antol et al. [5] which allows for indoor and outdoor scenes to be created. The clipart vocabulary consists of 20 deformable human models, 31 animals in various poses, and around 100 objects that are found in indoor (e.g., chair, table, sofa, fireplace, notebook, painting) and outdoor (e.g., sun, cloud, tree, grill, campfire, slide) scenes. The human models span different genders, races, and ages with 8 different expressions. They have limbs that are adjustable to allow for continuous pose variations. This combined with the large vocabulary of objects result in diverse scenes with rich semantics. Fig. 25 (*Top Row*) shows scenes that AMT workers created using this abstract scenes interface and vocabulary. Additional details, example scenes, and a sample of clipart objects are available on the project webpage.

### 3.1.2 Abstract Visual Humor (AVH) Dataset

This dataset consists of funny and unfunny scenes created by AMT workers, facilitating the study of visual humor at the scene level.

**Collecting Funny Scenes.** We collect 3.2K scenes via AMT by asking workers to create funny scenes that are meaningful, realistic, and that other people would also consider funny. This is to encourage workers to refrain from creating scenes with inside jokes or catering to a very personalized form of humor. A screenshot of the interface used to collect the data is available on the project webpage. We provide a

[1] `www.github.com/VT-vision-lab/abstract_scenes_v002`



(a) 0.1 (b) 1.5

(c) 4.0 (d) 4.0

**Figure 9:** Spectrum of scenes *(left to right)* in ascending order of funniness score, $F_i$ (Section 3.1.2) as rated by AMT workers.

random subset of the clipart vocabulary to each worker out of which at least 6 clipart objects are to be used to create a scene. In addition, we also ask the worker to give a brief description of why the scene is funny in a short phrase or sentence. We find that this encourages workers to be more thoughtful and detailed regarding the scene they create. Note that this is different from providing a caption to an image since this is a simple explanation of what the worker had in mind while creating the scene. Mining this data may be useful to better understand visual humor. However, in this work we focus on the harder task of understanding purely *visual* humor and do not use these explanations.

We also use an equal number (3.2K) of abstract scenes from [5] which are realistic, everyday scenes. We expect most of these scenes to be mundane (*i.e.*, not funny).

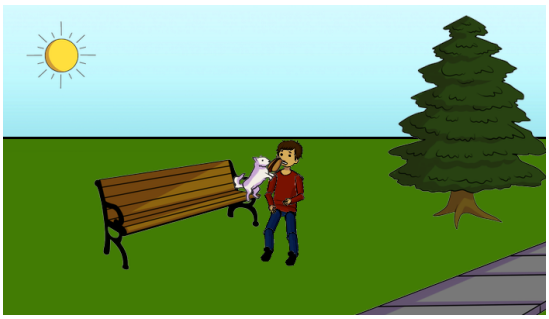
**Labeling Scene Funniness.** Anyone who has tried to be funny knows that humor is a subjective notion. A well-intending worker may create a scene that other people

do not find very funny. We obtain funniness ratings for each scene in the dataset from 10 different workers on AMT who do not see the creator's explanation of funniness. The ratings are on a scale of 1 to 5, where 1 is not funny and 5 is extremely funny. We define the *funniness score* of a scene, as the average of the 10 ratings for the scene. We found 10 ratings to be sufficient for good inter-human agreement. Further analysis is provided on the project webpage.

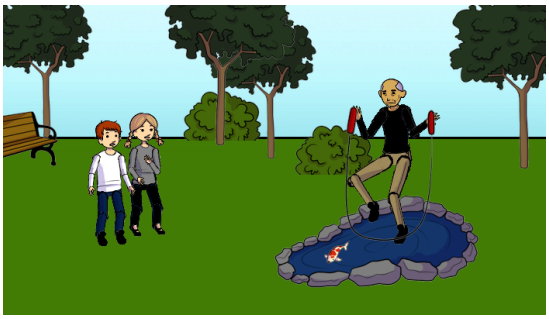
By plotting a distribution of these scores, we determine the optimal threshold that best separates scenes that were intended to be funny (*i.e.*, workers were specifically asked to create a funny scene) and other scenes (*i.e.*, everyday scenes from [5], where workers were not asked to create funny scenes). We label all scenes that have a funniness score greater than threshold as *funny* and all scenes with a lower funniness score as *unfunny*. This re-labeling results in 522 *unintentionally funny* scenes (i.e., scenes from [5], which were determined to be funny), and 682 *unintentionally unfunny* scenes (*i.e.*, well-intentioned worker outputs which were deemed not funny by the crowd). In total, this dataset contains 6,400 scenes (3,028 funny scenes and 3,372 unfunny scenes). We randomly split these scenes into train, val, and test sets having 60%, 20%, and 20% of the scenes, respectively. We refer to this dataset as the AVH dataset.

**Humor Techniques.** To better understand the different sources of humor in our dataset, we collect human annotations of the different techniques are used to depict humor in each scene. We create a list of humor techniques that are motivated by existing humor theories, based on patterns that we observe in funny scenes, and the audio-visual humor typology by Buijzen *et al*. [33]: *person doing something unusual*, *animal doing something unusual*, *clownish behavior (*i.e.*, goofiness)*, *too many objects*, *somebody getting hurt*, *somebody getting scared* and *somebody getting angry*.

We choose a subset of 200 funny scenes from the AVH dataset. We show each of these scenes to 10 different AMT workers and ask them to choose all the humor



(a) Animal doing something unusual



(b) Person doing something unusual

(c) Person getting hurt

(d) Person getting scared

**Figure 10:** Top voted scenes by humor technique (Section 3.1.2).

techniques that are depicted. Our options also included *none of the above reasons*, which also prompted workers to briefly explain what other unlisted technique depicted in the scene made it funny. However, we observe that this option was rarely used by workers. This may indicate that most of our scenes can be explained well by one of the listed humor techniques. Fig. 10 shows the top voted images corresponding to the 4 most popular techniques of humor. We find that the techniques that involve animate objects – *animal doing something unusual* and *person doing something unusual* are voted higher than any other technique by a large margin. For 75% of the scenes, at least 3 out of 10 workers picked one of these two techniques. We observe that this *unusualness* or *incongruity* is generally caused by objects occurring in an unusual context in the scene.

Introducing or eliminating incongruities can alter the funniness of a scene. An elderly person kicking a football while simultaneously skateboarding (Fig. 11, *bottom*)

is incongruous and hence considered funny. However, when the person is replaced by a young girl, this is is not incongruous and hence not funny. Such incongruities that can alter the funniness of a scene serves as our motivation to collect the Funny Object Replaced dataset which we describe next.

### 3.1.3 Funny Object Replaced (FOR) Dataset

Replacing objects in a scene is a technique to manipulate incongruities (and hence funniness) in a scene. For instance, we can change funny interactions (which are unexpected by our common sense) to interactions that are *normal* according to our mental model of the world. We use this technique to collect a dataset which consists of funny scenes and their altered unfunny counterparts. This enables the study of humor in a scene at the *object-level*.

We show funny scenes from the AVH dataset and ask AMT workers to make the least number of replacements in the scene to render the originally funny scene unfunny. The motivation behind this is to get a precise signal of which objects in the scene contribute to humor and what they can be replaced with to reduce/eliminate humor, while keeping the underlying structure of the scene the same. We ask workers to replace an object with another object that is as similar as possible to the first object and keep the scene realistic. This helps us understand fine-grained semantics that causes a specific object category to contribute to humor. There could be other ways to manipulate humor, *e.g.*, by adding, removing, or moving objects in a scene, *etc.* but in our work we employ only the technique of replacing objects. We find that this technique is very effective in altering the funniness of a scene. Our interface did not allow people to add, remove, or move the objects in the scene. A screenshot of the interface used to collect this dataset is available on the project webpage.

For each of the 3,028 funny scenes in the AVH dataset, we collect *object-replaced* scenes from 5 different workers resulting in 15,140 unfunny counterpart scenes. As

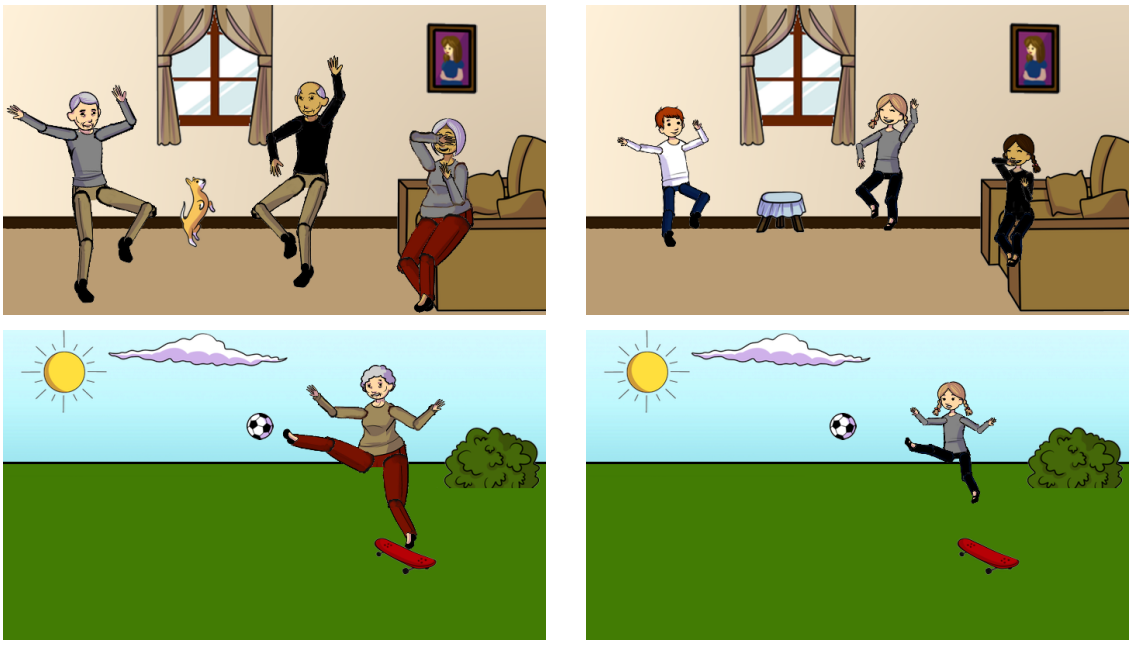


**Figure 11:** Funny scenes (*left*) and *one* among the 5 corresponding object-replaced unfunny counterparts (*right*) from the FOR dataset (see Section 3.1.3). For each funny scene, we collect an unfunny counterpart from a different worker.

a sanity check, we collect funniness ratings (via AMT) for 750 unfunny counterpart scenes. We observe that they indeed have an average funniness of 1.10, which is smaller than that of their corresponding original funny scenes (whose average funniness is 2.66). Fig. 11 shows two pairs of funny scenes and their object-replaced unfunny counterparts. We refer to this dataset as the FOR dataset.

Given the task posed to workers (altering a funny scene to make it unfunny), it is natural to use this dataset to train a model to reduce the humor in a scene. However, this dataset can also be used to train flipped models that can increase the humor in a scene as shown in Section 3.3.2.3.

## *3.2 Approach*

We propose and model two tasks that we believe demonstrate an understanding of some aspects of visual humor:

1. Predicting how funny a given scene is.

2. Altering the funniness of a scene.

The models that perform the above tasks are described in Section 3.2.2 and Section 3.2.3, respectively. The features used in the models are described first (Section 3.2.1).

### 3.2.1 Features

key advantage of using abstract scenes is that the scenes are trivially densely-annotated (via the scene creation interface), allowing us to compute rich semantic features. Recall that our abstract scenes are of two scene types (indoor and outdoor) and our vocabulary consists of 150 object categories including humans, animals, small objects, and large objects. We compute features at different scales, namely instance-level (each instantiation of an object type is treated separately) and scene-level.

Abstract scenes are trivially densely annotated which we use to compute rich semantic features. Recall that our interface allows two types of scenes (indoor and outdoor) and our vocabulary consists of 150 object categories. We compute both scene-level and instance-level features.

1. **Instance-Level Features**

(a) **Object embedding (150-d)** is a distributed representation that captures the context in which an object category usually occurs. We learn this representation using a word2vec-style continuous Bag-of-Words model [149]. The model tries to predict the presence of an object category in the scene, given the context provided by other instances of objects in the scene. Specifically, in a scene, 5 (randomly chosen) instances are projected onto a vector space. The sum of projections of the 5 object representations is used to predict the object category of the 6th instance. We train the single-layer (150-d) neural network [148] with multiple 6-item subsets of instances from each scene.

The network is trained using Stochastic Gradient Descent (SGD) with a momentum of 0.9. We use 11K scenes (that were not intended to be funny) from the dataset

collected in [5] to train the model. Thus, we learn representations of objects occurring in natural contexts which are not funny. A visualization of the object embeddings is available on the project webpage.

(b) **Local embedding (150-d)** For each instantiation of an object in the scene, we compute a weighted sum of object embeddings of all the other instances in the scene. The weight of every other instance is its inverse square-root distance with respect to the instance under consideration.

2. **Scene-Level Features**

(a) **Cardinality (150-d)** is a Bag-of-Words representation that indicates the number of instances of each object category that are present in the scene.

(b) **Location (300-d)** is a vector of the horizontal and vertical coordinates of every object in the scene. When multiple instances of an object category are present, we consider location of the instance closest to the center of the scene.

(c) **Scene Embedding (150-d)** is the sum of object embeddings of all objects present in the scene.

### 3.2.2 Predicting Funniness Score

We train a Support Vector Regressor (SVR) that predicts the funniness score, $F_i$ for a given scene $i$. The model regresses to the $F_i$ computed from ratings given by AMT workers (described in Section 3.1.2) on scenes from the AVH dataset (Section 3.1.2). We train the SVR on the scene-level features (described in Section 3.2.1) and perform an ablation study.

### 3.2.3 Altering Funniness of a Scene

We learn models to alter the funniness of a scene – from funny to unfunny and *vice versa*. Our two-stage pipeline involves:

1. Detecting objects that contribute to humor.
2. Identifying suitable replacement objects from 1. to make the scene unfunny (or funny), while keeping it realistic.

**Detecting Humor.** We train a multi-layer perceptron (MLP) on scenes from the FOR dataset to make a binary prediction on each object instance in the scene – whether it should be replaced to alter the funniness of a scene or not. The input is a 300-d vector formed by concatenating object embedding and local embedding features. The MLP has two hidden layers comprising of 300 and 100 units respectively, to which ReLU activation is applied. The final layer has 2 neurons and is used to perform binary classification (replace or not) using cross-entropy loss. We train the model using SGD with a base learning rate of 0.01 and momentum of 0.9. We also trained a model with skip-connections that considers the predictions made on other objects when making a prediction on a given object. However, this did not result in significant performance gains.

**Altering Humor.** We train an MLP to perform a 150-way classification to predict potential replacer objects (from the clipart vocabulary), given an object predicted to be replaced in a scene. The model's input is a 300-d vector formed by concatenating local embedding and object embedding features. The classifier has 3 hidden layers of 300 units each, with ReLU non-linearities. The output layer has 150 units over which we compute soft-max loss. We train the model using SGD with a base learning rate of 0.1, momentum of 0.9, and a dropout ratio of 0.5. The label for an instance is the index of the replacer object category used by the worker. Due to the large diversity of viable replacer objects that can alter humor in a scene, we also analyze the top-5 predictions of this model. We train two models – one on funny scenes, and another on their unfunny counterparts from the FOR dataset. Thus, we learn models to alter the funniness in a scene in one direction – funny to unfunny or vice versa. Although we could train the pipeline end-to-end, we train each stage separately so that we can

evaluate them separately and isolate their errors (for better interpretability).

## *3.3 Results*

We discuss the performance of our models in the two visual humor tasks of:

1. Predicting how funny a given scene is (Section 3.3.1)
2. Altering funniness of a scene (Section 3.3.2).

We discuss the quantitative results of our model in altering an unfunny scene to make it funny in Section 3.3.2.2), and the *vice versa* in Section 3.3.2.3. In Section 3.3.3, we report qualitative results through human studies.

### 3.3.1 Predicting Funniness Score

This section presents performance of the SVR (Section 3.2.2) that predicts the funniness score of a scene.

**Metric.** We use average relative error to quantify our model's performance computed as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{|Predicted\ F_i - Ground\ Truth\ F_i|}{Ground\ Truth\ F_i} \tag{1}$$

where $N$ is the number of test scenes and $F_i$ is the funniness score for the test scene $i$.

**Baseline:** The baseline model always predicts the average funniness score of the training scenes.

**Model.** As shown in Table 1, we observe that our model trained using combinations of different scene-level features (described in Sec. 3.2.1) performs better than the baseline model. We see that Location features perform slightly better than Cardinality. This makes sense because Location features also have occurrence information. The Embedding does not have location information and hence does worse. Due to some redundancy (all features have occurrence information), combining them does not improve performance.

**Table 1:** Performance of different feature combinations in predicting funniness score $F_i$ of a scene.

| **Features** | **Avg. Rel. Err.** |
|---|---|
| Avg. Prediction Baseline | 0.3151 |
| Embedding | 0.2516 |
| Cardinality | 0.2450 |
| Location | 0.2400 |
| Embedding + Cardinality + Location | 0.2400 |

### 3.3.2 Altering Funniness of a Scene

We discuss the performance in the tasks of identifying objects in a scene that contribute to humor (Section 3.2.2) and replacing those objects with other objects to reduce (or increase) humor (Section 3.2.3).

#### *3.3.2.1 Predicting Objects to be Replaced*

We train this model to detect objects instances that are funny in the scene. It makes a binary prediction whether each instance should be replaced or not.

**Metric.** Along with naïve accuracy (% of correct predictions, *i.e.*, Acc.), we also report average class-wise accuracy (*i.e.*, Avg. Cl. Acc.) to determine the performance of our model for this task. As the data is skewed, with the majority class being *not-replace*, we require our model to perform well both class-wise and as a whole.

**Baselines:**

1. **Priors.** We always predict that an instance should not be replaced. We also compute a stronger baseline that replaces an object if it is replaced at least T% of the time in training data. T was set to 20 based on the validation set.
2. **Anomaly Detection.** From the scene embedding, we subtract the object embedding of the object under consideration. We then compute the cosine similarity

of the resultant scene embedding with the object embedding. Objects with the least similarity with the scene are the anomalous objects in the scene. This is similar to finding the odd-one-out given a group of words [148]. Objects that have a cosine similarity less than a threshold T with the scene are predicted as anomalous objects and are replaced. A modification to this baseline is to replace K objects that are least similar to the scene. Based on performance on the validation set, T and K are determined to be 0.8 and 4, respectively.

**Model.** Table 2 compares the performance of our model with the baselines described above. We observe that the baseline based on priors performs better than anomaly detection. This is perhaps not surprising because the prior-based baseline, while naïve, is `supervised` in the sense that it relies on statistics from the training dataset of which objects tend to get replaced. On the other hand, anomaly detection is completely unsupervised since it only captures the context of objects in *normal* scenes. Our approach performs better than the baseline approaches in identifying objects that contribute to humor.

On average, we observe that our model replaces 3.67 objects for a given image as compared to an average of 2.54 objects replaced in the ground truth. This bias to replace more objects ensures that a given scene becomes significantly less funny than the original scene. We observe that the model learns that in general, animate objects like humans and animals are potentially stronger sources of humor compared to inanimate objects. It is interesting to note that the model also learns fine-grained detail, *e.g.*, to replace older people playing outdoors (which may be considered funny) with younger people (Fig. 12, top row).

#### *3.3.2.2 Making a Scene Unfunny*

Given that an object is predicted to be replaced in the scene, the model has to also predict a suitable replacer object. In this section, we discuss the performance of

**Table 2:** Performance of predicting whether an object should be replaced or not, for the task of altering a funny scene to make it unfunny. As the data is skewed with the majority class being `not-replace`, we require our model to perform well both class-wise and as a whole.

| Method | Avg. Cl. Acc. | Acc. |
|---|---|---|
| Priors (do not replace) | 50% % | **79.86**% |
| Priors (object's tendency to be replaced) | 73.13 % | 71.5% |
| Anomaly detection (threshold distance) | 62.16 % | 58.30% |
| Anomaly detection (top-K objects) | 63.01 % | 64.31% |
| Our model | **74.45%** | 74.74% |

the model in predicting these replacer objects. This model is trained and evaluated using ground truth annotations of objects that are replaced by humans in a scene. This helps us isolate performance between predicting *which objects to replace* and predicting *suitable replacers* .

**Metric.** In order to evaluate the performance of the model on the task of replacing funny objects in the scene to make it unfunny, we use the top-5 metric (similar to ImageNet [198]), *i.e.*, if any of our 5 most confident predictions match the ground truth, we consider that as a correct prediction.

**Baselines:**

1. **Priors.** Every object is replaced by one of its 5 most frequent replacers in the training set.
2. **Anomaly Detection.** We subtract the embedding of the object that is to be replaced from the scene embedding. The 5 objects from the clipart vocabulary that are most similar (in the embedding space) to this resultant scene embedding are the ones that contextually `fit in`.

**Model.** We observe that the performance trend in Table 3 is similar to that observed in the previous section (Sec. 3.3.2.1), i.e., our model performs better than priors, which performs better than anomaly detection. By qualitative inspection, we find that our

**Table 3:** Performance of predicting which object to replace with, for the task of altering a funny scene to make it unfunny.

| Method | Top-5 accuracy |
|---|---|
| Priors (top 5 GT replacers) | 24.53% |
| Anomaly detection (object that `fits` into scene) | 7.69% |
| Our model | **29.65%** |

top prediction is intelligent, but lazy. It eliminates humor in most scenes by choosing to replace objects contributing to humor with other objects that blend well into the background. By relegating an object to the background, it is rendered inactive and hence, cannot be contribute to humor in the scene. For e.g., the top prediction is frequently `plant` in indoor scenes and `butterfly` in outdoor scenes. The 2nd prediction is both intelligent and creative. It effectively reduces humor while also ensuring diversity of replacer objects. Subsequent predictions from the model tend to be less meaningful. Qualitatively, we find the 2nd most confident prediction to be the best compromise.

**Full pipeline.** Fig. 12 shows qualitative results from our full pipeline (predicting objects to replace and predicting their replacers) using the 2nd predictions made by our model.

#### *3.3.2.3 Making a Scene Funny*

We train our full pipeline model used in Sec. 3.3.2.2 on scenes from the FOR dataset to perform the task of altering an unfunny scene to make it funny. Some qualitative results are shown in Fig. 13.

### 3.3.3 Human Evaluation

We conducted two human studies to evaluate our full pipeline:

1. **Absolute:** We ask 10 workers to rate the funniness of the scene predicted by our model on a scale of 1-5. We then compare this with the $F_i$ of the input funny

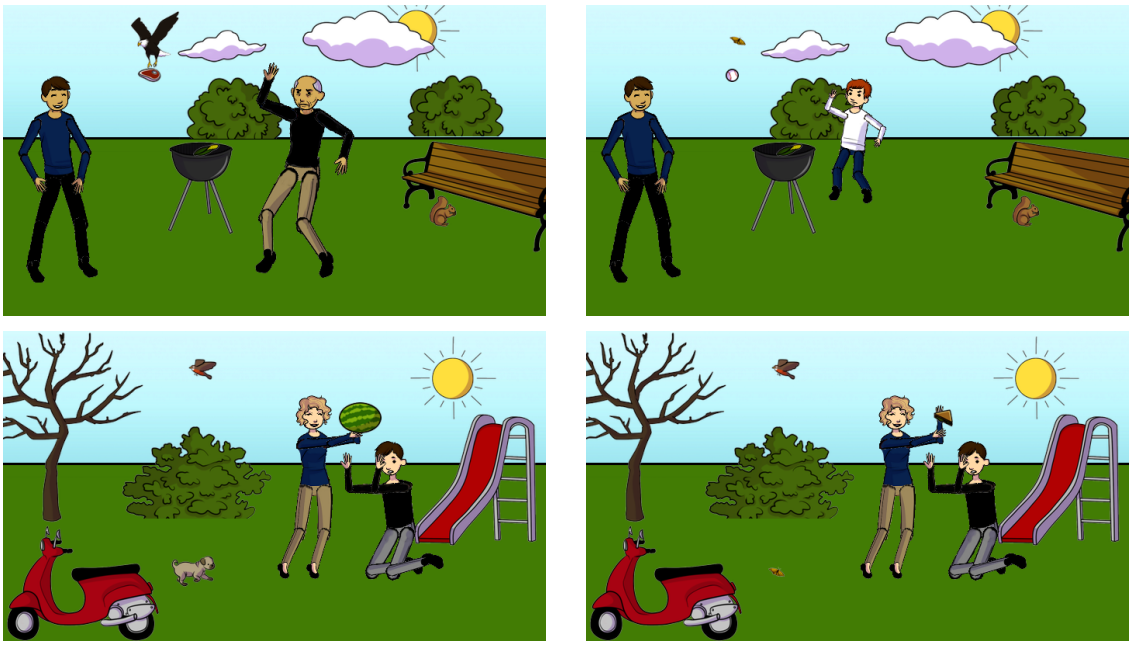


**Figure 12:** Fully automatic result of altering an input funny scene *(left)* into an unfunny scene *(right)*.

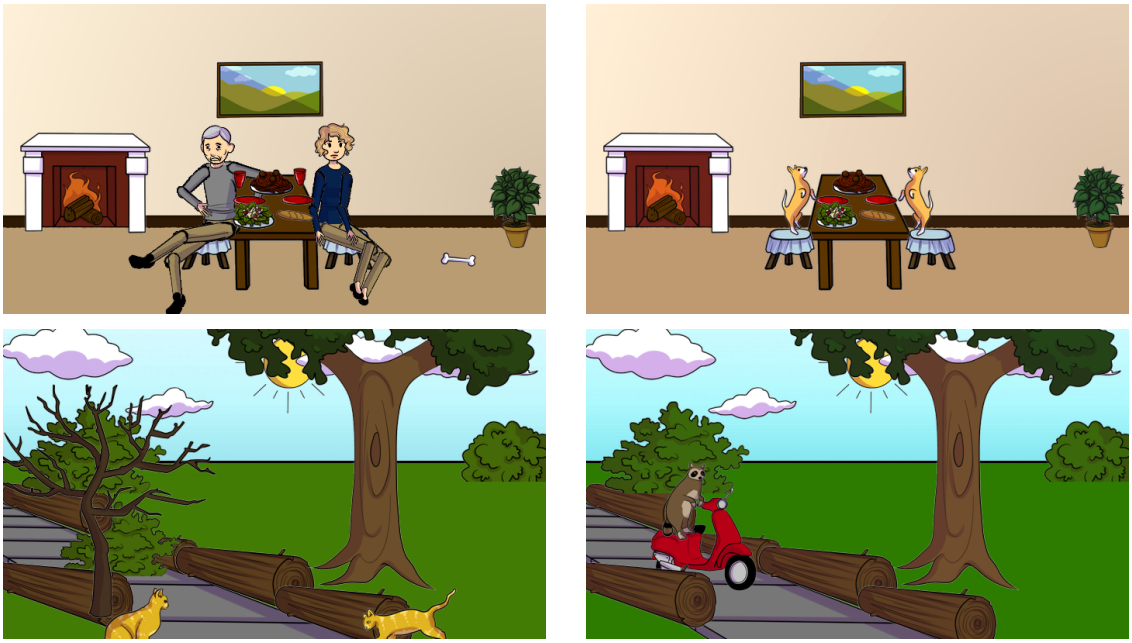


**Figure 13:** Fully automatic result of altering an input unfunny scene *(left)* into a funny scene *(right)*.

scene.

2. **Relative:** We show 5 workers the input scene and the predicted scene (in random order) and ask them to indicate which scene is funnier.

**Funny to unfunny.** As expected, the output scenes from our model are less funny than the input funny scenes on average. The average $F_i$ of the input funny test scenes is 2.69. This is 1.05 points higher than the output unfunny scenes whose average $F_i$ is 1.64. Unsurprisingly, in relative evaluation, workers find our output scenes to be less funny than the input funny scenes 95% of the time.

**Unfunny to funny.** During absolute evaluation, we find that the average $F_i$ of scenes made funny by our model is 2.14. This is a relatively high score, considering that the average $F_i$ score of the corresponding originally funny scenes that were created by workers is 2.69. Interestingly, the relative evaluation can be perceived as a *Turing test* of sorts, where we show workers the model's output funny scene and the original funny scene created by workers. 28% of the time, workers picked the model's scenes to be *funnier*.

## *3.4 Discussion*

Humor is a subtle and complex human behavior. It has many forms ranging from slapstick which has a simple physical nature, to satire which is nuanced and requires an understanding of social context [239]. Understanding the entire spectrum of humor is a challenging task. It demands perception of fine-grained differences between seemingly similar scenarios. *E.g.*, a teenager falling off his skateboard (such as in America's Funniest Home Videos[2]) could be considered funny but an old person falling down the stairs is typically horrifying. Due to these challenges some people even consider computational humor to be an 'AI-complete' problem [22, 98].

[2] `www.afv.com`

While understanding fine-grained semantics is important, it is interesting to note that there exists a qualitative difference in the way humor is perceived in abstract and real scenes. Since abstract scenes are not photorealistic, they afford us *suspension of reality*. Unlike real images, the content depicted in an abstract scene is benign. Thus, people are likely to find the depiction more funny [141]. In our everyday lives, we come across a significant amount of humorous content in the form of comics and cartoons to which our computational models of humor are directly applicable. They can also be applied to learn semantics that can extend to photorealistic images as demonstrated by Antol et al. [6].

Recognizing funniness involves violation of our mental model of how the world ought to be [145]. In verbal humor, the first few lines of the joke (set-up) build up the world model and the last line (punch line) goes against it. It is unclear what forms our mental model when we look at images. Is it our priors about the world around us formed from our past experiences? Is it because we attend to different regions of the image when we look at it and gradually build an expectation of what to see in the rest of the image? These are some interesting questions regarding visual humor that remain unanswered.

## *3.5 Conclusion*

In this Chapter, we took a step towards understanding and predicting visual humor. We collected two datasets of abstract scenes which enable the study of humor at different levels of granularity. We trained a model to predict the *funniness score* of a given scene. We also explored the different sources of humor depicted in the funny scenes via human studies. We trained models using incongruity-based humor to alter a scene's funniness. The models learned that in general, animate objects like humans and animals contribute more to humor compared to inanimate objects. Our model outperformed a strong anomaly detection baseline, demonstrating that

detecting humor involves something more than just anomaly detection. In human studies of the task of making an originally funny scene unfunny, humans found our model's output to be less funny 95% of the time. In the task of making a normal scene funny, our evaluation can be interpreted as a *Turing test* of sorts. Scenes made funny by our model were found to be funnier 28% of the time when compared with the original funny scenes created by workers. Note that our model would match humans at 50%. We hope that addressing the problem of studying visual humor using abstract scenes and the two datasets that are made public would stimulate further research in this new direction.

# CHAPTER IV

# MULTI-MODAL HUMOR

,

> "*Wit is the sudden marriage of ideas which before their union were not perceived to have any relation.*"
>
> – Mark Twain

Witty remarks are often contextual, i.e., grounded in a specific situation. Developing computational models that can emulate rich forms of interaction like contextual humor, is a crucial step towards making human-AI interaction more natural and more engaging [249]. E.g., witty chatbots could help relieve stress and increase user engagement by being more personable and human-like. Bots could automatically post witty comments (or suggest witty responses) on social media, chat, or messaging.

The absence of large scale corpora of witty captions and the prohibitive cost of collecting such a dataset (being witty is harder than just describing an image) makes the problem of producing contextually witty image descriptions challenging.

In this Chapter, we attempt to tackle the challenging task of producing witty (pun-based) remarks for a given (possibly boring) image. Our approach is inspired by a two-stage cognitive account of humor appreciation [215] which states that a perceiver experiences humor when a stimulus such as a joke, captioned cartoon, etc., causes an *incongruity*, which is shortly followed by *resolution*.

We introduce an incongruity in the perceiver's mind while describing an image by using an unexpected word that is phonetically similar (pun) to a concept related to the image. E.g., in Fig. 14(b), the expectations of a perceiver regarding the image (bear,

(a) **Generated**: a poll (pole) on a city street at night.
**Retrieved**: the light knight (night) chuckled.
**Human**: the knight (night) in shining armor drove away.

(b) **Generated**: a bare (bear) black bear walking through a forest.
**Retrieved**: another reporter is standing in a bare (bear) brown field.
**Human**: the bear killed the lion with its bare (bear) hands.

**Figure 14:** Sample images and witty descriptions from 2 models, and a human. The words inside '()' (e.g., pole and bear) are the puns associated with the image, i.e., the source of the unexpected puns used in the caption (e.g., poll and bare).

stones, etc.) is momentarily disconfirmed by the (phonetically similar) word 'bare'. This incongruity is resolved when the perceiver parses the entire image description. The incongruity followed by resolution can be perceived to be witty.[1]

We build two computational models based on this approach to produce witty descriptions for an image. First, a model that retrieves sentences containing a pun that are relevant to the image from a large corpus of stories [254]. Second, a model that generates witty descriptions for an image using a modified inference procedure during image captioning which includes the specified pun word in the description.

Our paper makes the following contributions: To the best of our knowledge, this is the first work that tackles the challenging problem of producing a witty natural language remark in an everyday (boring) context. We present two novel models to produce witty (pun-based) captions for a novel (likely boring) image. Our models rely on linguistic wordplay. They use an unexpected pun in an image description during inference/retrieval. Thus, they do not require to be trained with witty captions. Humans vote the descriptions from the top-ranked *generated* captions 'wittier' than three baseline approaches. Moreover, in a Turing test-style evaluation, our model's best image description is found to be wittier than a witty human-written caption[2] 55% of the time when the human is subject to the same constraints as the machine regarding word usage and style.

## *4.1 Approach*

**Extracting tags**. The first step in producing a contextually witty remark is to identify concepts that are relevant to the context (image). At times, these concepts are directly available as e.g., tags posted on social media. We consider the general case where such tags are unavailable, and automatically extract tags associated with

[1]Indeed, a perceiver may fail to appreciate wit if the process of 'solving' (resolution) is trivial (the joke is obvious) or too complex (they do not 'get' the joke).

[2]This data is available on the author's webpage.

an image.

We extract the top-5 object categories predicted by a state-of-the-art Inception-ResNet-v2 model [216] trained for image classification on ImageNet [56]. We also consider the words from a (boring) image description (generated from [226]). We combine the classifier object labels and words from the caption (ignoring stopwords) to produce a set of tags associated with an image, as shown in Fig. 15. We then identify concepts from this collection that can potentially induce wit.

**Identifying puns**. We attempt to induce an incongruity by using a pun in the image description. We identify candidate words for linguistic wordplay by comparing image tags against a list of puns.

We construct the list of puns by mining the web for differently spelled words that sound exactly the same (heterographic homophones). We increase coverage by also considering pairs of words with 0 edit-distance, according to a metric based on fine-grained articulatory representations (AR) of word pronunciations [105]. Our list of puns has a total of 1067 unique words (931 from the web and 136 from the AR-based model).

The pun list yields a set of puns that are associated with a given image and their phonologically identical counterparts, which together form the *pun vocabulary* for the image. We evaluate our approach on the subset of images that have non-empty pun vocabularies (about 2 in 5 images).

**Generating punny image captions**. We introduce an incongruity by forcing a vanilla image captioning model [226] to decode a *phonological counterpart* of a pun word associated with the image, at a specific time-step during inference (e.g., 'sell' or 'sighed', showed in orange in Fig. 15). We achieve this by limiting the vocabulary of the decoder at that time-step to only contain counterparts of image-puns. In following time-steps, the decoder generates new words conditioned on all previously decoded words. Thus, the decoder attempts to generate sentences that flow well based on

previously uttered words.

We train two models that decode an image description in forward (start to end) and reverse (end to start) directions, depicted as 'fRNN' and 'rRNN' in Fig. 15 respectively. The fRNN can decode words after accounting for the incongruity that occurs early in the sentence and the rRNN is able to decode the early words in the sentence after accounting for the incongruity that can occur later. The forward RNN and reverse RNN generate sentences in which the pun appears in each of the first $T$ and last $T$ positions, respectively.[3]

**Retrieving punny image captions**. As an alternative to our approach of *generating* witty remarks for the given image, we also attempt to leverage natural, human-written sentences which are relevant (yet unexpected) in the given context. Concretely, we retrieve natural language sentences[4] from a combination of the Book Corpus [254] and corpora from the NLTK toolkit [129]. The retrieved sentences each (a) contains an incongruity (pun) whose counterpart is associated with the image, and (b) has support in the image (contains an image tag). This yields a pool of candidate captions that are perfectly grammatical, a little unexpected, and somewhat relevant to the image (see Sec. 4.2).

**Ranking**. We rank captions in the candidate pools from both generation and retrieval models, according to their log-probability score under the image captioning model. We observe that the higher-ranked descriptions are more relevant to the image and grammatically correct. We then perform non-maximal suppression, i.e., eliminate captions that are similar[5] to a higher-ranked caption to reduce the pool to a smaller, more diverse set. We report results on the 3 top-ranked captions. We describe the

[3]For an image, we choose $T = \{1, 2, ..., 5\}$ and beam size $= 6$ for each decoder. This generates a pool of 5 (T) $*$ 6 (beam size) $*$ 2 (forward + reverse decoder) $= 60$ candidates.

[4]To prevent the context of the sentence from distracting the perceiver, we consider sentences with $< 15$ words. Overall, we are left with a corpus of about 13.5 million sentences.

[5]Two sentences are similar if the cosine similarity between the average of the Word2Vec [149] representations of words in each sentence is $\geq 0.8$.

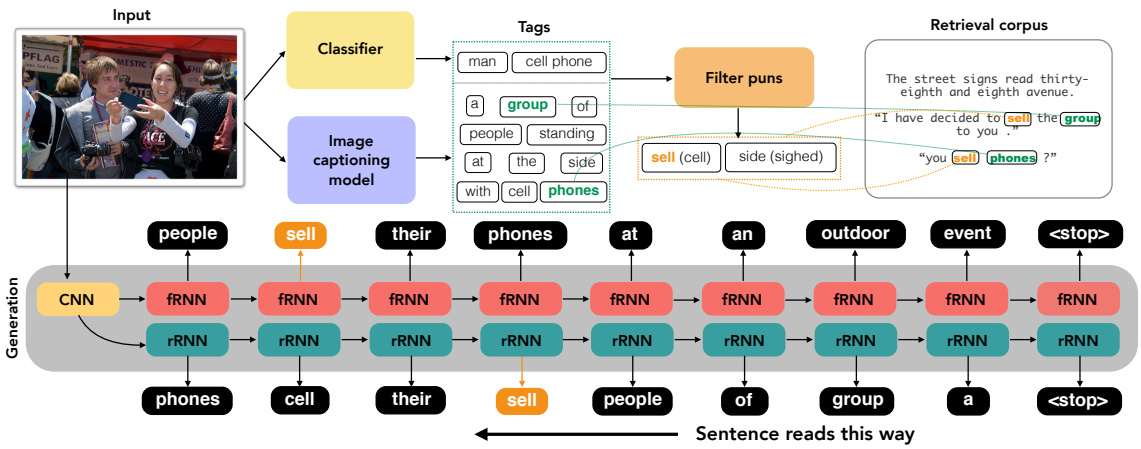



**Figure 15:** Our models for generating and retrieving image descriptions containing a pun (see Sec. 4.1).

effect of design choices in the appendix.

## *4.2 Results*

**Data**. We evaluate witty captions from our approach via human studies. 100 random images (having associated puns) are sampled from the validation set of COCO [125].

**Baselines**. We compare the wittiness of descriptions generated by our model against 3 qualitatively different baselines, and a human-written witty description of an image. Each of these evaluates a different component of our approach. **Regular inference** generates a fluent caption that is relevant to the image but is not attempting to be witty. **Witty mismatch** is a human-written witty caption, but for a different image from the one being evaluated. This baseline results in a caption that is intended to be witty, but does not attempt to be relevant to the image. **Ambiguous** is a 'punny' caption where a pun word in the boring (regular) caption is replaced by its counterpart. This caption is likely to contain content that is relevant to the image, *and* it contains a pun. However, the pun is not being used in a fluent manner.

We evaluate the **image-relevance** of the top witty caption by comparing against a boring machine caption and a random caption (see supplementary).

**Evaluating annotations**. Our task is to generate captions that a layperson might

find witty. To evaluate performance on this task, we ask people on Amazon Mechanical Turk (AMT) to vote for the wittier among the given pair of captions for an image. We collect annotations from 9 unique workers for each relative choice and take the majority vote as ground-truth. For each image, we compare each of the generated 3 top-ranked and 1 low-ranked caption against 3 baseline captions and 1 human-written witty caption.[6]

**Constrained human-written witty captions**. We evaluate the ability of humans and automatic methods to use the given context *and pun words* to produce a caption that is perceived as witty. We ask subjects on AMT to describe a given image in a witty manner. To prevent observable *structural* differences between machine and human-written captions, we ensure consistent pun vocabulary (utilization of pre-specified puns for a given image). We also ask people to avoid first person accounts or quote characters in the image.

**Metric**. 16, we report performance of the generation approach using the Recall@K metric. For $K = 1, 2, 3$, we plot the percentage of images for which at least one of the $K$ 'best' descriptions from our model outperformed another approach.

**Generated captions vs. baselines.** As we see in Fig. 16, the top generated image description (top-1G) is perceived as wittier compared to all baseline approaches more often than not (the vote is >50% at $K = 1$). We observe that as $K$ increases, the recall steadily increases, i.e., when we consider the top $K$ generated captions, increasingly often, humans find at least one of them to be wittier than captions produced by baseline approaches. People find the top-1G for a given image to be wittier than mismatched human-written image captions, about 95% of the time. The top-1G is also wittier than a naive approach that introduces ambiguity about 54.2% of the

[6]This results in a total of 4 (captions) $*2$ (generation + retrieval) $*4$ (baselines + human caption) = 32 comparisons of our approach against baselines. We also compare the wittiness of the 4 generated captions against the 4 retrieved captions (see supplementary) for an image (16 comparisons). In total, we perform 48 comparisons per image, for 100 images.

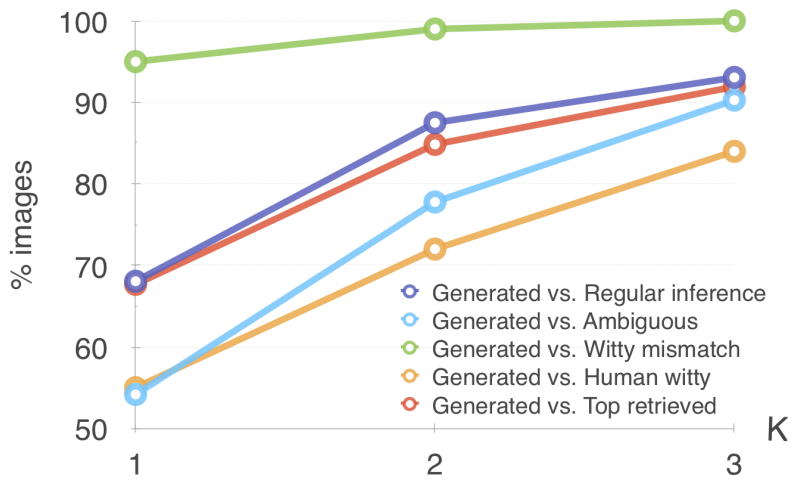



**Figure 16:** Wittiness of top-3 generated captions vs. other approaches. y-axis measures the % images for which at least one of $K$ captions from our approach is rated wittier than other approaches. Recall steadily increases with the number of generated captions ($K$).

time. When compared to a typical, boring caption, the generated captions are wittier 68% of the time. Further, in a head-to-head comparison, the generated captions are wittier than the retrieved captions 67.7% of the time. We also validate our choice of ranking captions based on the image captioning model score. We observe that a 'bad' caption, i.e., one ranked lower by our model, is significantly less witty than the top 3 output captions.

Surprisingly, when the human is constrained to use the same words and style as the model, the generated descriptions from the model are found to be wittier for 55% of the images. Note that in a Turing test, a machine would equal human performance at 50%[7]. This led us to speculate if the constraints placed on language and style might be restricting people's ability to be witty. We confirmed this by evaluating free-form human captions.

**Free-form Human-written Witty Captions.** We ask people on AMT to describe

[7]Recall that this compares how a witty description is constructed, given the image *and* specific pun words. A Turing test-style evaluation that compares the overall wittiness of a machine and a human would refrain from constraining the human in any way.

(a) **Generated**: a female tennis player caught (court) in mid swing.
**Retrieved**: i caught (court) thieves on the roof top.
**Human**: the man made a loud bawl (ball) when she threw the ball.

(b) **Generated**: a loop (loupe) of flowers in a glass vase.
**Retrieved**: the flour (flower) inside teemed with worms.
**Human**: piece required for peace (piece).

(c) **Generated**: a woman sell (cell) her cell phone in a city.
**Retrieved**: Wright (right) slammed down the phone.
**Human**: a woman sighed (side) as she regretted the sell.

(d) **Generated**: a bear that is bare (bear) in the water.
**Retrieved**: water glistened off her bare (bear) breast.
**Human**: you won't hear a creak (creek) when the bear is feasting.

**Figure 17:** Some qualitative examples from our approach. The top row contains selected examples of human-written witty captions, and witty captions generated and retrieved from our models. The examples in the bottom row are randomly picked.


(a) **Generated**: a loop (loupe) of scissors and a pair of scissors.
**Retrieved**: i continued slicing my pear (pair) on the cutting board.
**Human**: the scissors were near, but not clothes (close).


(b) **Generated**: a bored (board) bench sits in front of a window.
**Retrieved**: Wedge sits on the bench opposite Berry, bored (board).
**Human**: could you please make your pleas (please)!

(c) **Generated**: a bored (board) living room with a large window.
**Retrieved**: anya sat on the couch, feeling bored (board).
**Human**: the sealing (ceiling) on the envelope resembled that in the ceiling.

(d) **Generated**: a parking meter with rode (road) in the background.
**Retrieved**: smoke speaker sighed (side).
**Human**: a nitting of color didn't make the poll (pole) less black.

**Figure 18:** A few more qualitative examples from our approaches. The top row contains selected examples of human-written witty captions, and witty captions generated and retrieved from our models. The examples in the bottom row are randomly picked.

an image (using any vocabulary) in a manner that would be perceived as funny. As expected, when compared against automatic captions from our approach, human evaluators find free-form human captions to be wittier about 90% of the time compared to 45% in the case of constrained human witty captions. Clearly, human-level creative language with unconstrained sentence length, style, choice of puns, etc., makes a significant difference in the wittiness of a description. In contrast, our automatic approach is constrained by caption-like language, length, and a word-based pun list. Training models to intelligently navigate this creative freedom is an exciting open challenge.

**Qualitative analysis**. The generated witty captions exhibit interesting features like alliteration ('a bare black bear ...') in Fig. 14(b) and 17(c). At times, both the original pun (pole) and its counterpart (poll) make sense for the image (Fig. 14(a)). Occasionally, a pun is naively replaced by its counterpart (Fig. 18(b)) or rare puns are used (Fig. 17(b)). On the other hand, some descriptions (Fig. 18(a) and 18(d)) that are forced to utilize puns do not make sense. See supplementary for analysis of retrieval model.

The retrieved witty descriptions are retrieved from story-based corpora. They often contain sentences that describe a very specific situation or instance. Although these sentences are grounded in objects that are also present in the image, the entire sentence often contains a few words that are irrelevant for a given image, as we see in Fig. 18(b), Fig. 17(b) and Fig. 18(c). This is a likely reason for why a retrieved sentence containing a pun is perceived as less witty when compared with witty descriptions generated for the image.

## 4.3 Discussion

Since wit involves unexpectedness, the objective of describing an image in a witty manner often results in a trade-off between the description being witty and the description being relevant to the image. It may be interesting to study how the perceived wittiness of an image description varies as it includes more creative elements and becomes less relevant to the image. Another interesting factor that can influence perceived humor is presentation. For instance, the text in cartoons and memes are funny in their characteristic, informal font but may seem boring in other, more 'serious' font.

Producing a description for an image that is perceived as witty is challenging because the description must achieve the fine balance between lending itself to easy resolution by the perceiver while not being impossible or too trivial. There are other challenges, however. For instance, automatic image recognition and captioning models, despite the great strides of advancement in recent times, are still imperfect. In our approach, these are cascading sources of error which could adversely affect the perceived wittiness of an image caption.

In the work described above, we only consider the use of *words* that are perfect puns. Future work can extend our approach to explore the use of phrase-based and imperfect puns to create alternate interpretations of a sentence.

Our approach has no constraints on the modality of the input stimulus. It can be extended to generate witty responses to input stimuli of different modalities, ranging from abstract (cartoon-like) scenes, to describing a given video in a witty manner or to generate witty response (dialog) to a text input.

## 4.4 Conclusion

Wit is a form of rich interaction that is often grounded in a specific situation (e.g., a comment in response to an event). In this Chapter, we developed computational

models that can produce witty descriptions for a given image. Inspired by a cognitive account of humor appreciation, we employ linguistic wordplay, specifically puns, in image descriptions. We developed two approaches that involve retrieving witty descriptions for a given image from a large corpus of sentences, or generating them via an encoder-decoder neural network architecture. We compared our approach against meaningful baseline approaches via human studies and showed substantial improvements. We found that when a human is subject to similar constraints as the model regarding word usage and style, people vote the image descriptions generated by our model to be slightly wittier than human-written witty descriptions. Unsurprisingly, humans are almost always wittier than the model when they are free to choose the vocabulary, style, etc.

# CHAPTER V

# NARRATIVE

Sequencing is a task for children that is aimed at improving understanding of the temporal occurrence of a sequence of events in a narrative. The task is, given a jumbled set of images (and maybe captions) that belong to a single story, sort them into the correct order so that they form a coherent story. Our motivation in this work is to enable AI systems to better understand and predict the temporal nature of events in the world. To this end, we train machine learning models to perform the task of "sequencing".

Temporal reasoning has a number of applications such as multi-document summarization of multiple sources of, say, news information where the relative order of events can be useful to accurately merge information in a temporally consistent manner. In question answering tasks [192, 68, 236, 189], answering questions related to when an event occurs, or what events occurred prior to a particular event require temporal reasoning. A good temporal model of events in everyday life, i.e., a 'temporal common sense', could also improve the quality of communication between AI systems and humans.

Stories are a form of narrative sequences that have an inherent temporal common sense structure. We propose the use of visual stories depicting personal events to learn temporal common sense. We use stories from the Sequential Image Narrative Dataset (SIND) [97] in which a set of 5 aligned image-caption pairs together form a coherent story. Given an input story that is jumbled (Fig. 25(a)), we train machine learning models to sort them into a coherent story (Fig. 25(b)).[1]

[1]Note that 'jumbled' here refers to the loss of temporal ordering; image-caption pairs are still aligned.

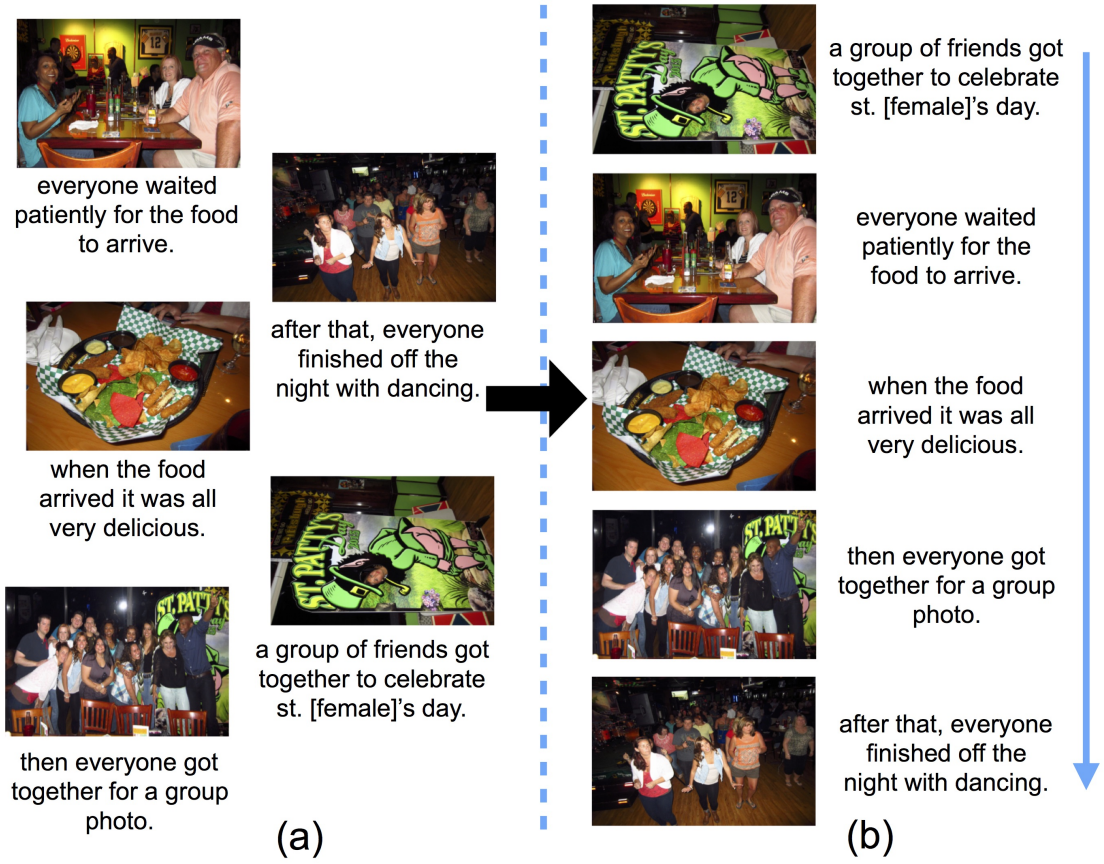



**Figure 19:** (a) The input is a jumbled set of aligned image-caption pairs. (b) Actual output of the system – an ordered sequence of image-caption pairs that form a coherent story.

Our contributions are as follows:

– We propose the task of visual story sequencing.

– We implement two approaches to solve the task: one based on individual story elements to predict position, and the other based on pairwise story elements to predict relative order of story elements. We also combine these approaches in a voting scheme that outperforms the individual methods.

– As features, we represent a story element as both text-based features from the caption and image-based features, and show that they provide complementary improvements. For text-based features, we use both sentence context and relative order based distributed representations.

– We show qualitative examples of our models learning temporal common sense.

## *5.1 Approach*

In this section, we first describe the two components in our approach: unary scores that do not use context, and pairwise scores that encode relative orderings of elements. Next, we describe how we combine these scores through a voting scheme.

### 5.1.1 Unary Models

Let $\sigma \in \Sigma_n$ denote a permutation of $n$ elements (image-caption pairs). We use $\sigma_i$ to denote the position of element $i$ in the permutation $\sigma$. A unary score $S_u(\sigma)$ captures the appropriateness of each story element $i$ in position $\sigma_i$:

$$S_u(\sigma) = \sum_{i=1}^{n} P(\sigma_i|i) \tag{2}$$

where $P(\sigma_i|i)$ denotes the probability of the element $i$ being present in position $\sigma_i$, which is the output from an $n$-way softmax layer in a deep neural network. We experiment with 2 networks –

(1) A language-alone unary model (Skip-Thought+MLP) that uses a Gated Recurrent Unit (GRU) proposed by Cho et al. [47] to embed a caption into a vector space. We use

the Skip-Thought [118] GRU, which is trained on the BookCorpus [254] to predict the context (preceding and following sentences) of a given sentence. These embeddings are fed as input into a Multi-Layer Perceptron (MLP).

(2) A language+vision unary model (Skip-Thought+CNN+MLP) that embeds the caption as above and embeds the image via a Convolutional Neural Network (CNN). We use the activations from the penultimate layer of the 19-layer VGG-net [210], which have been shown to generalize well. Both embeddings are concatenated and fed as input to an MLP.

In both cases, the best ordering of the story elements (optimal permutation) $\sigma^* = \text{argmax}_{\sigma \in \Sigma_n} S_u(\sigma)$ can be found efficiently in $O(n^3)$ time with the Hungarian algorithm [162]. Since these unary scores are not influenced by other elements in the story, they capture the semantics and linguistic structures associated with specific positions of stories *e.g.*, the beginning, the middle, and the end.

### 5.1.2 Pairwise Models

Similar to learning to rank approaches [89], we develop pairwise scoring models that given a pair of elements $(i, j)$, learn to assign a score:

$S([\![\sigma_i < \sigma_j]\!] \mid i, j)$ indicating whether element $i$ should be placed before element $j$ in the permutation $\sigma$. Here, $[\![\cdot]\!]$ indicates the Iverson bracket (which is 1 if the input argument is true and 0 otherwise). We develop and experiment with the following 3 pairwise models:

(1) A language-alone pairwise model (Skip-Thought+MLP) that takes as input a pair of Skip-Thought embeddings and trains an MLP (with hinge-loss) that outputs $S([\![\sigma_i < \sigma_j]\!] \mid i, j)$, the score for placing $i$ before $j$.

(2) A language+vision pairwise model (Skip-Thought+CNN+MLP) that concatenates the Skip-Thought and CNN embeddings for $i$ and $j$ and trains a similar MLP as above.

(3) A language-alone neural position embedding (NPE) model. Instead of using frozen Skip-Thought embeddings, we learn a task-aware ordered distributed embedding for sentences. Specifically, each sentence in the story is embedded $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n),\ \mathbf{x}_i \in \mathbb{R}^d_+$, via an LSTM [94] with ReLU non-linearities. Similar to the max-margin loss that is applied to negative examples by Vendrov et al. [224], we use an asymmetric penalty that encourages sentences appearing early in the story to be placed closer to the origin than sentences appearing later in the story.

$$\begin{aligned} L_{ij} &= \Big\| \max(0, \alpha - (\mathbf{x}_j - \mathbf{x}_i)) \Big\|^2 \\ Loss &= \sum_{1<=i<j=n} L_{ij} \end{aligned} \tag{3}$$

At train time, the parameters of the LSTM are learned end-to-end to minimize this asymmetric ordered loss (as measured over the gold-standard sequences). At test time, we use $S([\![\sigma_i < \sigma_j]\!] \mid i, j) = L_{ij}$. Thus, as we move away from the origin in the embedding space, we traverse through the sentences in a story. Each of these three pairwise approaches assigns a score $S(\sigma_i, \sigma_j | i, j)$ to an ordered pair of elements (i,j), which is used to construct a pairwise scoring model:

$$S_p(\sigma) = \sum_{1<=i<j<=n} \Big\{ S([\![\sigma_i < \sigma_j]\!]) - S([\![\sigma_j < \sigma_i]\!]) \Big\}, \tag{4}$$

by summing over the scores for all possible ordered pairs in the permutation. This pairwise score captures local contextual information in stories. Finding the best permutation $\sigma^* = \text{argmax}_{\sigma \in \Sigma_n} S_p(\sigma)$ under this pairwise model is NP-hard so approximations will be required. In our experiments, we study short sequences ($n = 5$), where the space of permutations is easily enumerable ($5! = 120$). For longer sequences, we can utilize integer programming methods or well-studied spectral relaxations for this problem.

### 5.1.3 Voting-based Ensemble

To combine the complementary information captured by the unary ($S_u$) and pairwise models ($S_p$), we use a voting-based ensemble. For each method in the ensemble, we

find the top three permutations. Each of these permutations ($\sigma^k$) then vote for a particular element to be placed at a particular position. Let $V$ be a vote matrix such that $V_{ij}$ stores the number of votes for $i^{th}$ element to occur at $j^{th}$ position, i.e., $V_{ij} = \sum_k [\![\sigma_i^k == j]\!]$). We use the Hungarian algorithm to find the optimal permutation that maximizes the votes assigned, i.e., $\sigma^*_{\text{vote}} = \text{argmax}_{\sigma \in \Sigma_n} \sum_{i=1}^{n} \sum_{j=1}^{n} V_{ij} \cdot [\![\sigma_i == j]\!]$. We experimented with a number of model voting combinations and found the combination of pairwise Skip-Thought+CNN+MLP and neural position embeddings to work best (based on a validation set).

## *5.2 Experiments*

### 5.2.1 Data

We train and evaluate our model on personal multi-modal stories from the SIND (Sequential Image Narrative Dataset) [97], where each story is a sequence of 5 images and corresponding story-like captions. The narrative captions in this dataset, e.g., "friends having a good time" (as opposed to "people sitting next to each other") capture a sequential, conversational language, which is characteristic of stories. We use 40,155 stories for training, 4990 for validation and 5055 stories for testing.

### 5.2.2 Metrics

We evaluate the performance of our model at correctly ordering a jumbled set of story elements using the following 3 metrics:

**Spearman's rank correlation** (Sp.) [213] measures if the ranking of story elements in the predicted and ground truth orders are monotonically related (higher is better).

**Pairwise accuracy** (Pairw.) measures the fraction of pairs of elements whose predicted relative ordering is the same as the ground truth order (higher is better).

**Average Distance** (Dist.) measures the average change in position of all elements in the predicted story from their respective positions in the ground truth story (lower is better).

**Table 4:** Performance of our different models and features at the sequencing task.

| Method | Features | Sp. | Pairw. | Dist. |
|---|---|---|---|---|
| Random Order | | 0.000 | 0.500 | 1.601 |
| Unary | SkipThought | 0.508 | 0.718 | 1.373 |
| | SkipThought + Image | 0.532 | 0.729 | 1.352 |
| Pairwise | SkipThought | 0.546 | 0.732 | 0.923 |
| | SkipThought + Image | 0.565 | 0.740 | 0.897 |
| Pairwise Order | NPE | 0.480 | 0.704 | 1.010 |
| Voting | SkipThought + Image (Pairwise) + NPE | **0.675** | **0.799** | **0.724** |

### 5.2.3 Results

**Pairwise Models vs Unary Models** As shown in Table 4, the pairwise models based on Skip-Thought features outperform the unary models in our task. However, the Pairwise Order Model performs worse than the unary Skip-Thought model, suggesting that the Skip-Thought features, which encode context of a sentence, also provide a crucial signal for temporal ordering of story sentences.

**Contribution of Image Features** Augmenting the text features with image features results in a visible performance improvement of both the model trained with unary features and the model trained with pairwise features. While image features by themselves result in poor performance on this task, they seem to capture temporal information that is complementary to the text features.

**Ensemble Voting** To exploit the fact that unary and pairwise models, as well as text and image features, capture different aspects of the story, we combine them using a voting ensemble. Based on the validation set, we found that combining the Pairwise Order model and the Pairwise model with both Skip-Thought and Image (CNN) features performs the best. This voting based method achieves the best performance


(a) First Position


(b) Second Position


(c) Third Position


(d) Fourth Position


(e) Fifth Position

**Figure 20:** Word cloud corresponding to most discriminative words for each position.

on all three metrics. This shows that our different approaches indeed capture complementary information regarding feasible orderings of caption-image pairs to form a coherent story.

For comparison to existing related work, we tried to duplicate the pipelined approach of Modi et al. [158]. For this, we first parse our story sentences to extract SVO (subject, verb, object) tuples (using the Stanford Parser [43]). However, this

step succeeds for only 60% of our test data. Now even if we consider a *perfect* downstream algorithm that *always* makes the correct position prediction given SVO tuples, the overall performance is still a Spearman correlation of just 0.473, i.e., the *upper bound* performance of this pipelined approach is *lower* than the performance of our text-only end-to-end model (correlation of 0.546) in Table 4.

**Confusion Matrix for Predicting Position of an Element** . We visualize the 5-way classification confusion matrix for our best performing method i.e., Voting ensemble of Pairwise Skip-Thought+Image(CNN) and Pairwise Order (Neural Position Embedding (NPE)) in Fig. 21. The block-diagonal matrix structure shows that the model predicts the first and the last element of a story reasonably well but is often confused by elements in the middle of the story. This visualization suggests that the model has learnt the *three act structure* in stories, i.e., the setup, the middle and the climax.

### 5.2.4 Qualitative Analysis

Visualizations of position predictions from our model demonstrate that it has learnt the *three act structure* [220] in stories – the setup, the middle and the climax. We also present success and failure examples of our sorting model's predictions.

**Predicted Stories** We present qualitative examples of story orders predicted by the best performing model in Fig. 5.2.4 and Fig. 5.2.4. Fig. 5.2.4 shows example stories in which the position of all elements are predicted correctly. Fig. 5.2.4 shows stories in which none of the positions are predicted correctly by our model. These two examples show that our model clearly fails when there is no inherent temporal order in the story either via language or images.

**Visualizing Temporal Common Sense** We visualize our model's *temporal common sense*, in Fig. 20. The word clouds show *discriminative words* – the words that the model believes are indicative of sentence positions in a story. The size of a word is proportional to the ratio of its frequency of occurring in that position to other positions. Some words like 'party', 'wedding', etc., probably because our model believes that the start the story describes the setup – the occasion or event. People often tend to describe meeting friends or family members which probably results in the discriminative words such as 'people', 'friend', 'everyone' in the second and the third sentences. Moreover, the model believes that people tend to conclude the stories using words like 'finally', 'afterwards', tend to talk about 'great day', group 'pictures' with everyone, etc.

In the word cloud in Fig. 24, we visualize the words that the model finds *discriminative* in correct predictions. These are words from *correctly* predicted stories that the model believes are indicative of sentence positions in a story. The size of a word is proportional to the ratio of its frequency of occurring in that position to other positions. Our model captures events such as 'carnival', 'reunion', and sports topics like 'baseball', 'soccer', 'skate' in the first position. This could be the case because the first sentence of a story usually introduces the event that the story is based on. In Fig. 24(e) (word-cloud of the last sentence), we also observe that the model correctly learns cue-words such as 'overall', and 'lastly'. It also learns words and events that frequently conclude stories such as 'returned', 'tired', 'winning', 'winner', and 'celebration'.

## *5.3 Conclusion*

In this Chapter, we proposed the task of "sequencing" in a set of image-caption pairs, with the motivation of learning temporal common sense. We implemented multiple neural network models based on individual and pairwise element-based predictions

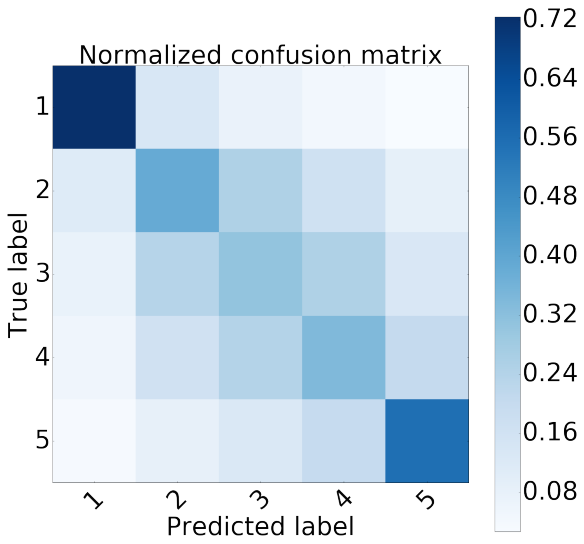



**Figure 21:** Confusion matrix for predictions from the best performing model i.e Voting ensemble of Pairwise Skip-Thought+image(CNN) and Pairwise Order Neural Position Embedding (NPE).

(and their ensemble), and utilize both image and text features, to achieve strong performance on the task. Our best system, on average, predicts the ordering of sentences to within a distance error of 0.8 (out of 5) positions. We also analyzed our predictions and show qualitative examples that demonstrate temporal common sense.


**Figure 22:** Examples of stories for which the temporal sequence of elements was predicted perfectly.

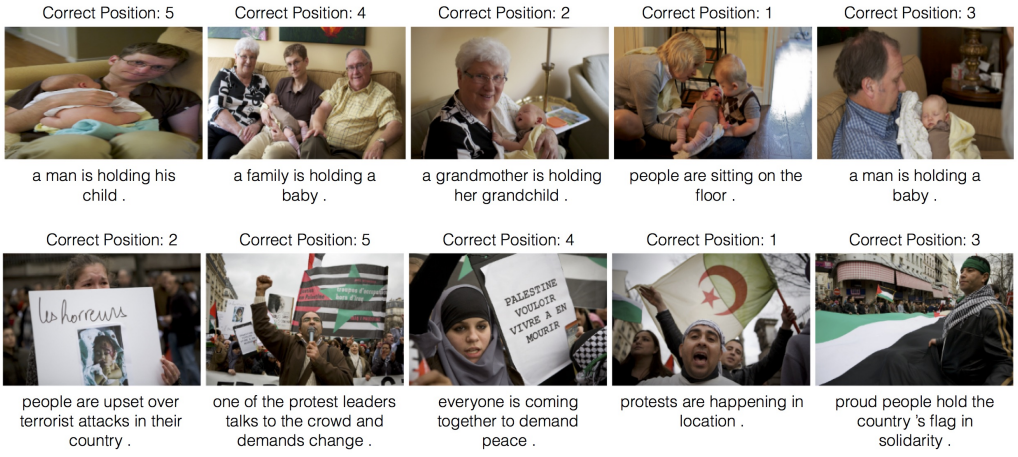
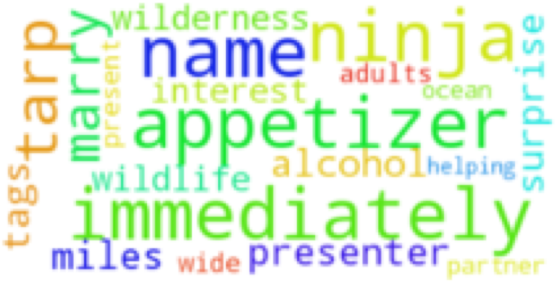
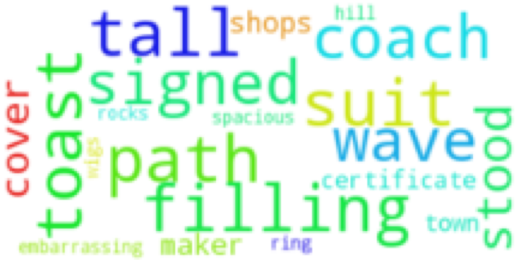



**Figure 23:** Examples of success and failure cases of temporal order prediction of story elements by our best performing model.

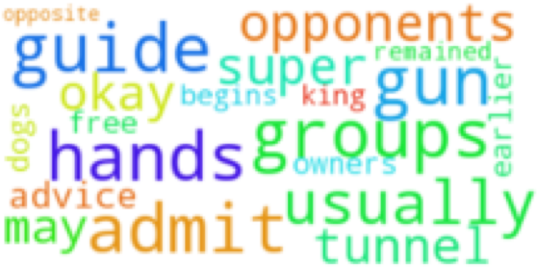

(a) First Position



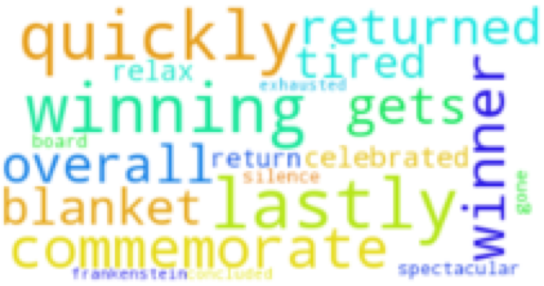

(b) Second Position




(c) Third position




(d) Fourth position




(e) Fifth position

**Figure 24:** Discriminative words in each position of all correctly predicted stories.

# CHAPTER VI

# PREDICTABILITY

As technology progresses, we are increasingly collaborating with AI agents in *interactive* scenarios where humans and AI work together as a team, e.g., in AI-assisted diagnosis, autonomous driving, etc. Thus far, AI research has typically only focused on the AI in such an interaction – for it to be more accurate, be more human-like, understand our intentions, beliefs, contexts, and mental states.

In this work, we argue that for human-AI interactions to be more effective, humans must also understand the AI's beliefs, knowledge, and quirks.

Many recent works generate human-interpretable 'explanations' regarding a model's decisions. These are usually evaluated offline based on whether human judges found them to be 'good' or to improve trust in the model. However, their contribution in an interactive setting remains unclear. In this work, we evaluate the role of explanations towards making a model predictable to a human.

We consider an AI trained to perform the multi-modal task of Visual Question Answering (VQA) [135, 5], i.e., answering free-form natural language questions about images. VQA is applicable to scenarios where humans actively elicit information from visual data, and naturally lends itself to human-AI interactions. We consider two tasks that demonstrate the degree to which a human understands their AI teammate (we call Vicki) – Failure Prediction (FP) and Knowledge Prediction (KP). In FP, we ask subjects on Amazon Mechanical Turk to predict if Vicki will correctly answer a given question about an image. In KP, subjects predict Vicki's exact response.

We aid humans in forming a mental model of Vicki by (1) familiarizing them with its behavior in a 'training' phase and (2) exposing them to its internal states via

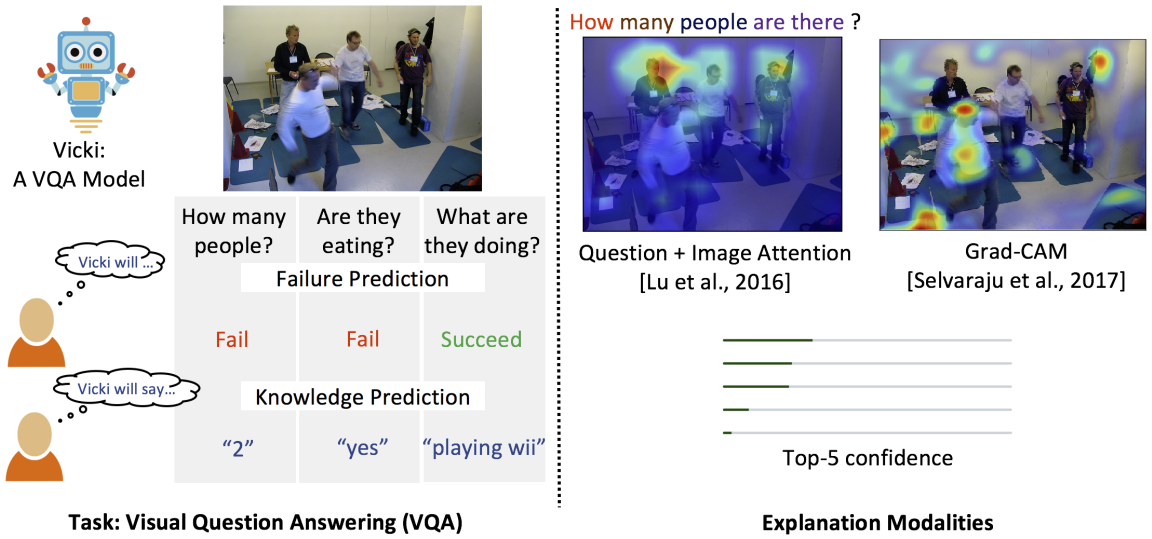



**Figure 25:** We evaluate the extent to which explanation modalities (right) and familiarization with a VQA model help humans predict its behavior – its responses, successes, and failures (left).

various explanation modalities. We then measure their FP and KP performance.

Our key findings are that (1) humans are indeed capable of predicting successes, failures, and outputs of the VQA model better than chance, (2) explicitly training humans to familiarize themselves with the model improves their performance, and (3) existing explanation modalities do not enhance human performance.

## *6.1 Setup*

**Agent.** We use the VQA model by Lu et al. [131] as our AI agent (that we call Vicki). The model processes the question at multiple levels of granularity (words, phrases, entire question) and at each level, has explicit attention mechanisms on both the image and the question[1]. It is trained on the train split of the VQA-1.0 dataset [5]. Given an image and a question about the image, it outputs a probability distribution over 1000 answers. Importantly, the model's image and question attention maps provide access to its 'internal states' while making a prediction.

[1]We use question-level attention maps in our experiments.

Vicky is *quirky* at times, i.e., has biases, albeit in a predictable way. Agrawal et al. [1] outlines several such quirks. For instance, Vicki has a limited capability to understand the image – when asked the color of a small object in the scene, say a soda can, it may simply respond with the most dominant color in the scene. Indeed, it may answer similarly even if no soda can is present, i.e. if the question is irrelevant.

Further, Vicki has a limited capability to understand free-form natural language, and in many cases, answers questions based only on the first few words of the question. It is also generally poor at answering questions requiring "common sense" reasoning. Moreover, being a discriminative model, Vicki has a limited vocabulary (1k) of answers. Additionally, the VQA 1.0 dataset contains label biases; therefore, the model is very likely to answer "white' to a "what color" question [84].

To get a sense for this, see Fig. 26 which depicts a clear pattern. In top-left, even when there is no grass, Vicki tends to latch on to one of the dominant colors in the image. For top-right, even when there are no people in the image, it seems to respond with what people could *plausibly* do in the scene if they were present. In this work, we measure to what extent lay people can pick up on these quirks by interacting with the agent, and whether existing explanation modalities help do so.

**Tasks. Failure Prediction (FP).** Given an image and a question about the image, we measure how well a person can predict if Vicki will successfully answer the question. A person can presumably predict the failure modes of Vicki well if they have a good sense of its strengths and weaknesses.

**Knowledge Prediction (KP).** In this task, we aim to obtain a fine-grained measure of a person's understanding of Vicki's behavior. Given a QI–pair, a subject guesses Vicki's exact response from a set of its output labels. Snapshots of our interfaces can be seen in Fig. 27.

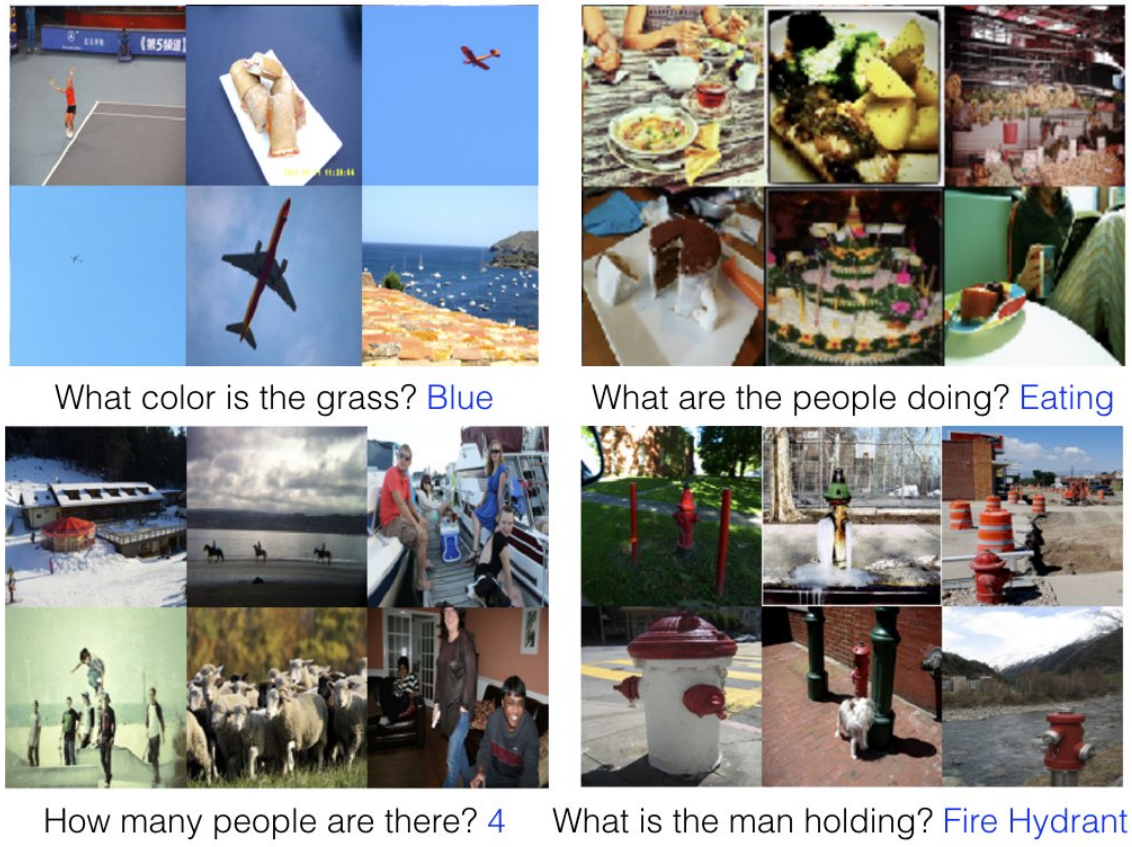



**Figure 26:** These montages highlight some of Vicki's quirks. For a given question, Vicki has the same response to each image in a montage. Common visual patterns (that Vicki presumably picks up on) within each montage are evident.

## *6.2 Experimental Setup*

In this section we investigate ways to make Vicki's behavior more predictable to a subject. We approach this by – providing instant feedback about Vicki's actual behavior on each QI pair once the subject responds, and exposing subjects to various explanation modalities that reveal Vicki's internal states before they respond.

**Data.** We identify a subset of questions in the VQA-1.0 [5] validation split that occur more than 100 times. We select 7 diverse questions[2] from this subset that are representative of the different types of questions (counting, yes/no, color, scene

[2]What kind of animal is this? What time is it? What are the people doing? Is it raining? What room is this? How many people are there? What color is the umbrella?

layout, activity, etc.) in the dataset. For each of the 7 questions, we sample a set of 100 images. For FP, the 100 images are random samples from the set of images on which the question was asked in VQA-1.0 val. For the KP task, these 100 images are random images from VQA-1.0 val. Ray et al. [183] found that randomly pairing an image with a question in the VQA-1.0 dataset results in about 79% of pairs being irrelevant. This combination of relevant and irrelevant QI-pairs allows us to test subjects' ability to develop a robust understanding of Vicki's behavior across a wide variety of inputs.

**Study setup.** We conduct our studies on Amazon Mechanical Turk. Each task (HIT) comprises of 100 QI-pairs where for simplicity (for the subject), a single question is asked across all 100 images. The annotation task is broken down into a train and test phase of 50 QI-pairs each. Over all settings, 280 workers took part in our study (1 unique worker per HIT), resulting in 28k human responses. Subjects were paid an average of $3 base plus $0.44 performance bonus, per HIT.

There are some challenges involved in scaling data-collection in this setting: (1) Due to the presence of separate train and test phases, our AMT tasks tend to be unusually long (mean HIT durations across the tasks of FP and KP = $10.11 \pm 1.09$ and $24.49 \pm 1.85$ min., respectively). Crucially, this also reduces the subject pool to only those willing to participate in long tasks. (2) Once a subject participates in a task, they cannot do another because their familiarity with Vicki would leak over. This constraint causes our analyses to require as many subjects as tasks. Since work division in crowdsourcing tasks follows a Pareto principle [127], this makes data collection very slow.

In light of these challenges, we focus on a small set of questions to systematically evaluate the role of training and exposure to Vicki's internal states.

### 6.2.1 Evaluating the role of familiarization

To familiarize subjects with Vicki, we provide them with instant feedback during the train phase. Immediately after a subject responds to a QI–pair, we show them whether Vicki actually answered the question correctly or not (in FP) or what Vicki's response was (in KP), along with a running score of how well they are doing. Once training is complete, no further feedback is provided and subjects are asked to make predictions for the test phase. At the end, they are shown their score and paid a bonus proportional to the score.

**Failure Prediction.** In FP, always guessing that Vicki answers 'correctly' results in 58.29% accuracy, while subjects do slightly better and achieve 62.66% accuracy, even without prior familiarity with Vicki (No Instant Feedback (IF)). Further, we find that subjects that receive training via instant feedback (IF) achieve 13.09% higher mean accuracies than those who do not (see Fig 25; IF vs No IF for FP (left)).

**Knowledge Prediction.** In KP, answering each question with Vicki's most popular answer overall ('no') would lead to an accuracy of 13.4%. Additionally, answering each question with its most popular answer *for that question* leads to an accuracy of 31.43%. Interestingly, subjects who are unfamiliar with Vicki (No IF) achieve 21.27% accuracy – better than the most popular answer overall, but worse than the question-specific prior over its answers. The latter is understandable as subjects unfamiliar with Vicki do not know which of its 1000 possible answers the model is most likely to predict for each question.

We find that mean performance in KP with IF is 51.11%, 29.84% higher than KP without IF (see Fig 25; IF vs No IF for KP (right)). It is apparent that just from a few (50) training examples, subjects succeed in building a mental model of Vicki's behavior that generalizes to new images. Additionally, the 29.84% improvement over No IF for KP is significantly larger than that for FP (13.09%). This is understandable because a priori (No IF), KP is a much harder task as compared to FP due to the

increased space of possible subject responses given a QI-pair, and the combination of relevant and irrelevant QI-pairs in the test phase.

Questions such as 'Is it raining?' have strong language priors – to these Vicki often defaults to the most popular answer ('no'), irrespective of image. On such questions, subjects perform considerably better in KP once they develop a sense for Vicki's inherent biases via instant feedback. For open-ended questions like 'What time is it?', feedback helps subjects (1) narrow down the 1000 potential options to the subset that Vicki typically answers with – in this case time periods such as 'daytime' rather than actual clock times and (2) identify correlations between visual patterns and Vicki's answer. In other cases like 'How many people are in the image?' the space of possible answers is clear a priori, but after IF subjects realize that Vicki is bad at detailed counting and bases its predictions on coarse signals of the scene layout.

### 6.2.2 Evaluating the role of explanations

In this setting, we show subjects an image, a question, and one of the explanation modalities described below. We experiment with 3 qualitatively different modalities (see Fig.25, right):

**Confidence of top-5 predictions.** We show subjects Vicki's confidence in its top-5 answer predictions from its vocabulary as a bar plot (of course, we do not show the actual top-5 predictions).

**Attention maps.** Along with the image we show subjects the spatial attention map over the image and words of the question which indicate the regions that Vicki is looking at and listening to, respectively.

**Grad-CAM.** We use the CNN visualization technique by Selvaraju et al. [203], using the (implicit) attention maps corresponding to Vicki's most confident answer.

**Automatic approaches**. We also evaluate automatic approaches to detect Vicki's

failure from its internal states. We find that both, a decision stump on Vicki's confidence in its top answer, and on the entropy of its softmax output, result in an FP accuracy of 60% on our test set. A Multi-layer Perceptron (MLP) trained on Vicki's output 1000-way softmax to predict success vs failure, achieves an FP accuracy of 81%. Training on just top-5 softmax outputs achieves an FP accuracy of 61.43%. Training an MLP which takes as input question features (average word2vec embeddings [150] of words in the question) concatenated with image features (fc7 from VGG-19) to predict success vs failure (which we call ALERT following [251]) achieves an FP accuracy of 65%. Training an MLP on identical question features as above but concatenated with Grad-CAM saliency maps leads to FP accuracy of 73.14%. [3] Note that we only report machine results to put human accuracies in perspective. We do not draw any inferences about the relative capabilities of both.

**Results.** Average performance of subjects in the test phases of FP and KP, for different experimental settings are summarized in Fig. 25. In the first setting, we show subjects an explanation modality with instant feedback (IF+Explanation). For reference, also see performance of subjects provided with IF and no explanation modality (IF).

We observe that on both FP and KP, subjects who received an explanation along with IF show no statistically significant difference in performance compared to those who did not. We see in Fig. 25, that both bootstrap based standard error (95% confidence intervals) overlap significantly.

Seeing that explanations in addition to IF does not outperform an IF baseline, we next measure whether explanations help a user not already familiar with Vicki via IF. That is, we evaluate if explanations help against a No IF baseline by providing an explanation only in the *test* phase, and no IF (see Fig 25; No IF + Explanation). Additionally, we also experiment with providing IF and an explanation *only* during

[3] These methods are trained on 66% of VQA-1.0 val. The remaining data is used for validation.

the train phase (see Fig 25; IF + Explanation (Train Only)), to measure whether access to internal states during training can help subjects build better intuitions for model behavior without needing access to internal states at test time. In both settings however, we observe no statistically significant difference in performance over the No IF and IF baselines, respectively.[4]

**Explanations help in narrow domain**. In the above section, we observe that explanations don't seem to improve predictability of a model consistently for all input samples across the board. However, prior research on explanation modalities has shown that they do help improve predictability of a model in certain specific scenarios. For instance, Selvaraju et al. [203] perform an experiment where they evaluate how class-discriminative their explanation from a classification model is. Specifically, they extract Grad-CAM visualizations for *each of the two object classes* in an image. They then show each visualization along with the input image to a human subject and ask them to identify the object that is highlighted in the image. Human subjects correctly identify the object in the image, about 61% of the time. This result obtained on a classification model, encourages us to utilize explanations to improve predictability of a VQA model.

Along these lines, a colleague in my lab performed experiments to identify if explanations from Grad-CAM do indeed improve predictability on a narrow domain of input questions and images. Specifically, they chose to study explanations for the question, 'what animal is this?' that was asked on images containing two animal categories. They identified 70 such images. The human subjects were shown the question, the image and asked to guess the response from the model. Another set of subjects were shown the GradCam explanations in addition to the input image and question, and asked to predict the response. Each input instance was shown to 9

[4]When piloting the tasks ourselves, we found it easy to 'overfit' to the explanations and hallucinate patterns.

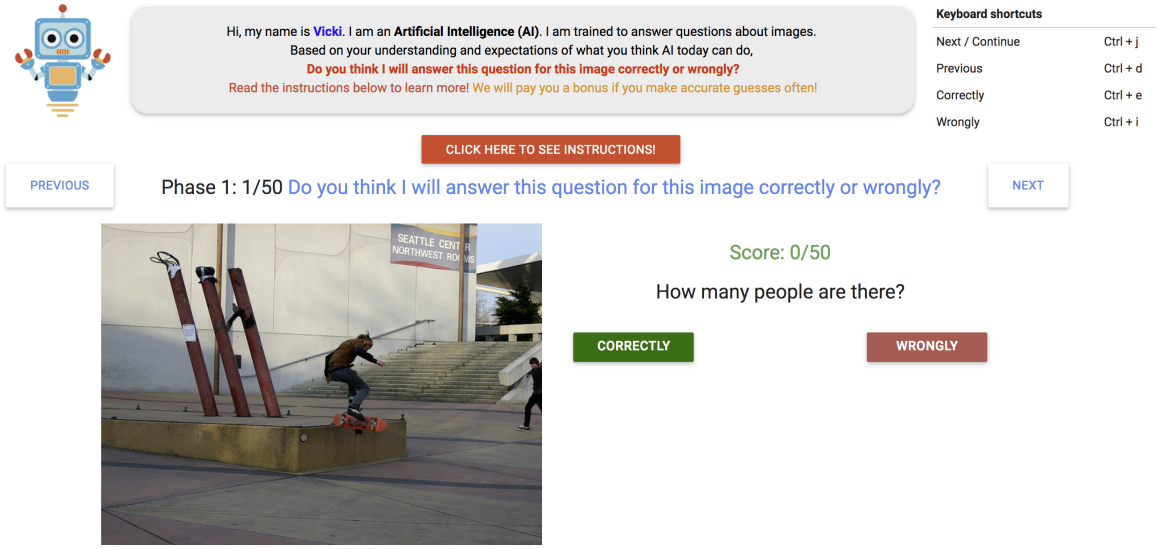
subjects.

They found that when human subjects were shown the input image and question, they accurately guessed the output from the model, about 67% of the time. When they were shown explanations from Grad-CAM attention maps in addition to the input image and question, their accuracy increased to 80%. This result clearly establishes the utility of explanations for a narrow domain of input to a VQA model. Understanding the largest scope for which explanations are useful, is a challenging, open research problem.

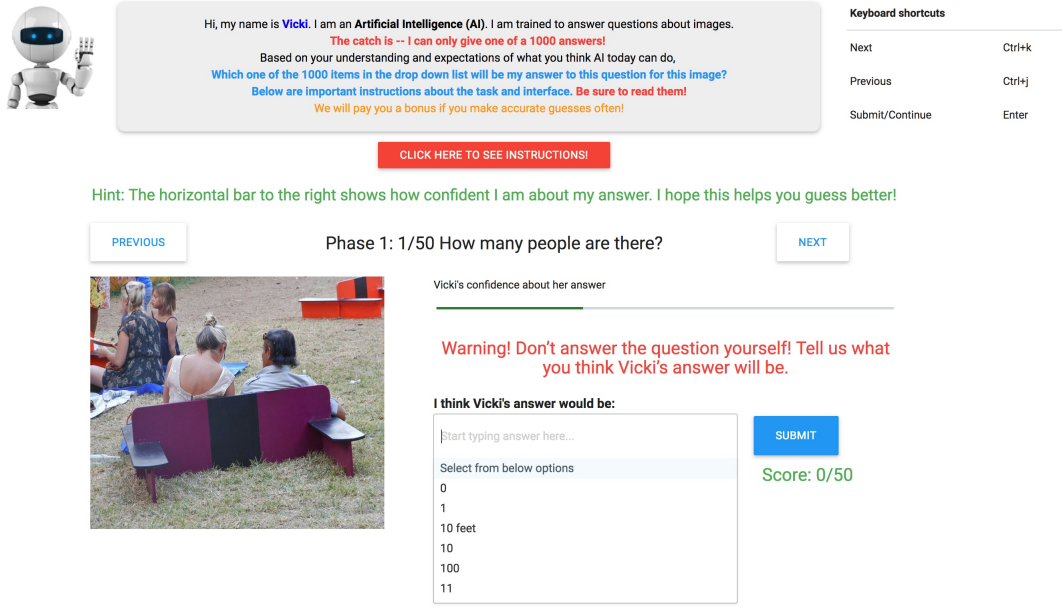
## *6.3 Conclusion*

As technology progresses, human-AI teams are inevitable. In this Chapter, we argued that for these teams to be more effective, it is essential to improve the humans' understanding of the strengths, weaknesses, quirks, and tendencies of AI. We instantiated these ideas in the domain of Visual Question Answering (VQA), by proposing two tasks that help measure how well a human 'understands' a VQA model (we call Vicki) – Failure Prediction (FP) and Knowledge Prediction (KP). We found that lay people indeed get better at predicting Vicki's behavior using just a few 'training' examples, but surprisingly, existing popular explanation modalities do not help make its failures or responses more predictable.

Relevant to the increased interest in interpretable AI, this work presented a novel framework for evaluation of explanation modalities that is grounded in specific tasks (FP and KP). Developing explanation modalities which explicitly aid humans in building a better model of the AI's behavior in a collaborative setting is an interesting direction for future work. Future work involves closing the loop and evaluating the extent to which improved human performance at FP and KP translates to improved success of human-AI teams at accomplishing a shared goal. In the next chapter, we present a co-operative human-AI game that is a natural fit for such an evaluation.




(a) The Failure Prediction (FP) interface.




(b) The Knowledge Prediction (KP) interface.

**Figure 27:** (a) A person guesses if a VQA model (Vicki) will answer this question for this image correctly or wrongly. (b) A person guesses what Vicki's exact answer will be for this QI–pair.

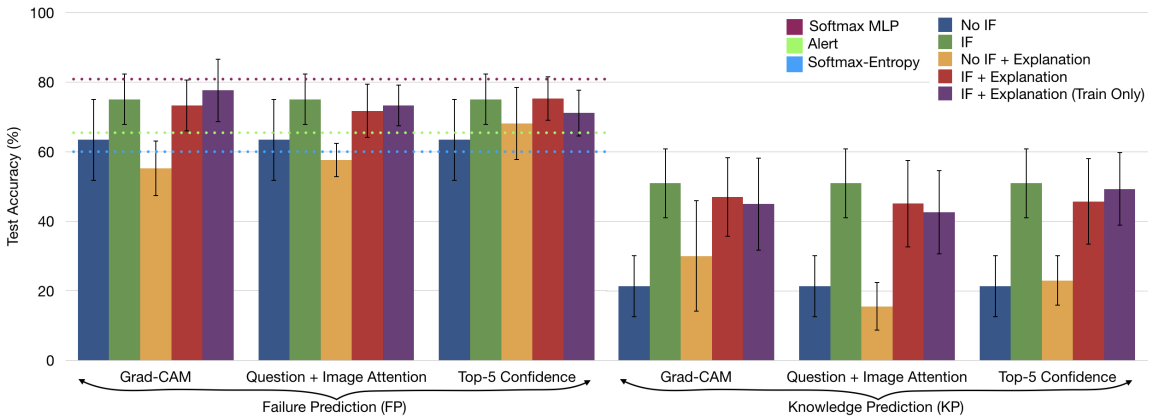



**Figure 28:** Average performance across subjects for Failure Prediction and Knowledge Prediction, across different settings: with or without (1) Instant feedback (IF) in the train phase, and (2) an explanation modality. Explanation modalities are shown in both train and test phases unless stated otherwise. Error bars are 95% confidence intervals from 1000 bootstrap samples. Note that the dotted lines are various machine approaches applied to FP.

# CHAPTER VII

# EXPLANATIONS IN HUMAN-AI TEAMS

Humans are now collaborating with AI across applications ranging from medical diagnosis, driving vehicles, to scheduling meetings. Moreover, the increasing acceptance of the use of AI makes it likely that human-AI teams will soon become widespread. While the performance of AI (specifically, deep neural networks) is rapidly improving on a number of tasks as evaluated by quantitative metrics, it is unclear to what extent the performance of these models can be leveraged by a human collaborating with it.

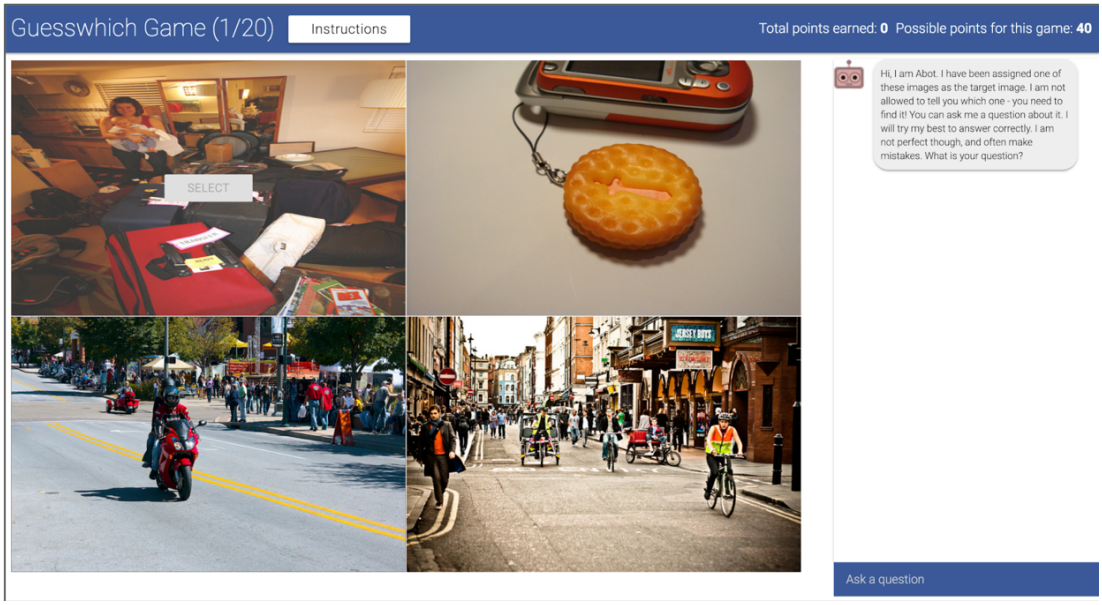
Existing research on human-agent teams [142] suggests that higher transparency of the agent leads to more effective performance of the human-agent team. We hypothesize that in the context of human-AI teams, the team could be more effective if the AI is more predictable to a human. In Chapter 6, we evaluated the predictability of a model to a lay person by asking them to predict the behavior of a model, given inputs to the model (and optionally explanations). While useful, this measure is only a proxy for the actual performance of a lay person collaborating with the model on a downstream, goal-driven task.

In this work, we propose the use of an interactive, collaborative, human-AI game of GuessWhich as a way to evaluate the human-AI team in a goal-driven co-operative task. GuessWhich, which was introduced in our earlier work [41], is a game that is similar to the game of '20 questions'. In GuessWhich, a human subject attempts to guess the 'secret' image known to the AI, from a pool of $N$ images. The human quizzes the model to gain more information regarding the secret image, and makes a final guess regarding the image (see details in Sec. 7.1.1). We add a feature to this interface that allows human subjects to also view the explanations from the model

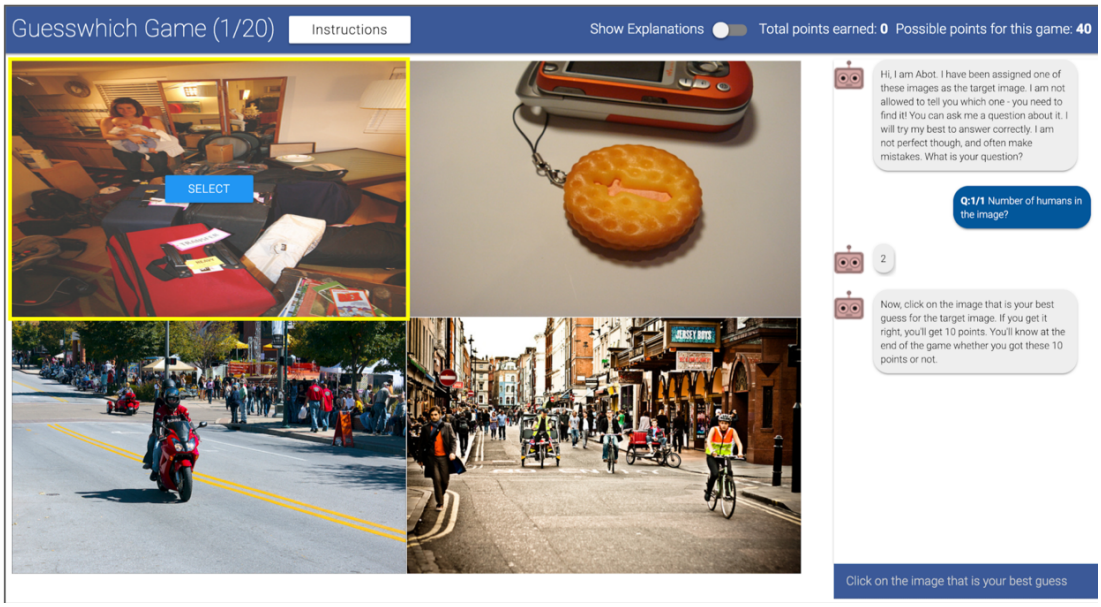
regarding its answer to the subject's question. Screenshots of the interface without explanations are shown in Fig. 29, Fig. 31 and with explanations are shown in Fig. 30, Fig. 32 respectively.

We hypothesize that an explanation that is consistent with the model's answer for the secret image, would bias the human subject towards picking the secret image. We perform experiments with a saliency based visual explanation [203], and text explanations [243] that provide a rationale for the VQA model's answer. We evaluate the performance of the human-AI team before and after the subject has access to the model's explanations. An increase in performance would demonstrate the utility of explanations from the model in the context of a collaborative human-AI team.

In the following sections, we first describe the gameplay in detail, followed by a brief description of the AI (VQA models), the explanation modalities (Grad-CAM and text explanations), and the human subjects. We then describe the experimental settings and present our findings from the data, followed by analysis. Finally, we discuss a few interesting aspects of the task and future research directions.

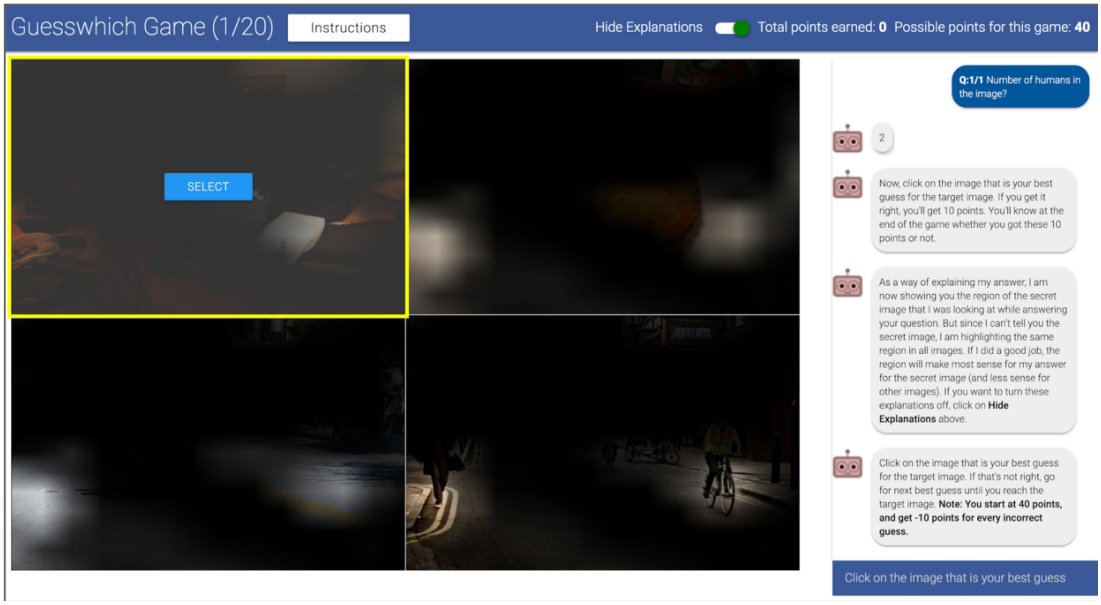



(a) Initial interface before subject's interaction.

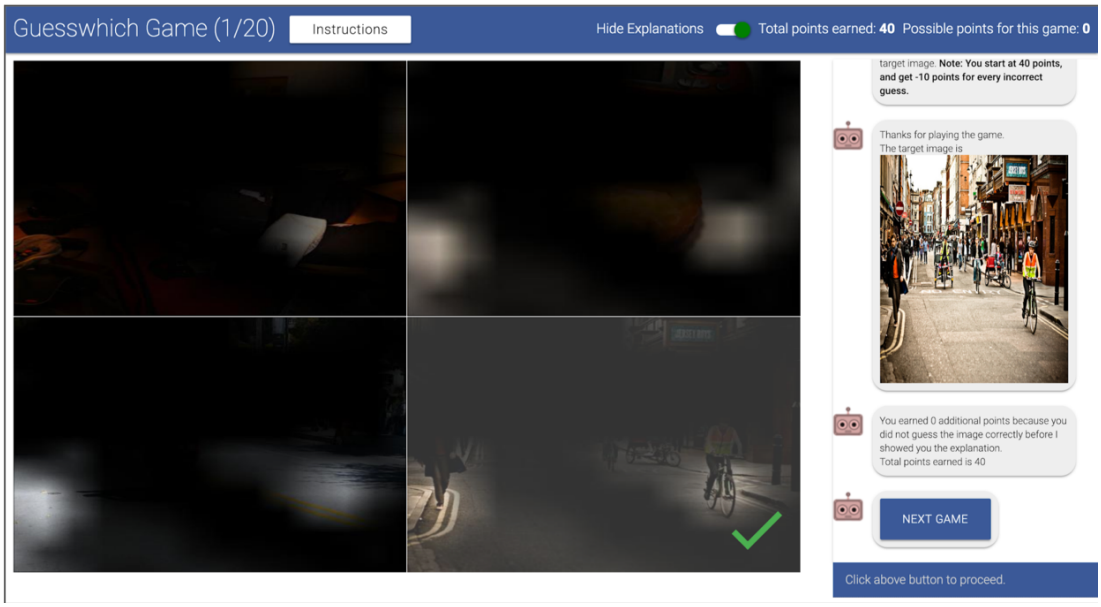



(b) Subject clicks on an image based on the model's response to their question.

**Figure 29:** Screenshots of the game interface without explanations. The subject is first shown the pool of images. The subject asks a question to the model ('Number of humans in the image?') based on the pool of images. The model responds to the subject's question ('2'). The subject then selects an image as a guess for the secret image.




(a) Grad-CAM explanation from the model is overlaid on all images.




(b) Subject makes a guess based on the model's response and the explanation.

**Figure 30:** Screenshots of the game interface with Grad-CAM explanations. Once the subject selects an image as their guess, the Grad-CAM explanation from the model (heat-maps) is overlaid on each of the images. Based on the Grad-CAM heat-map, i.e., the "model's explanation of where in the image it was looking while answering the question", the user then selects an image as a guess for the secret image.

## *7.1 Goal-driven task*

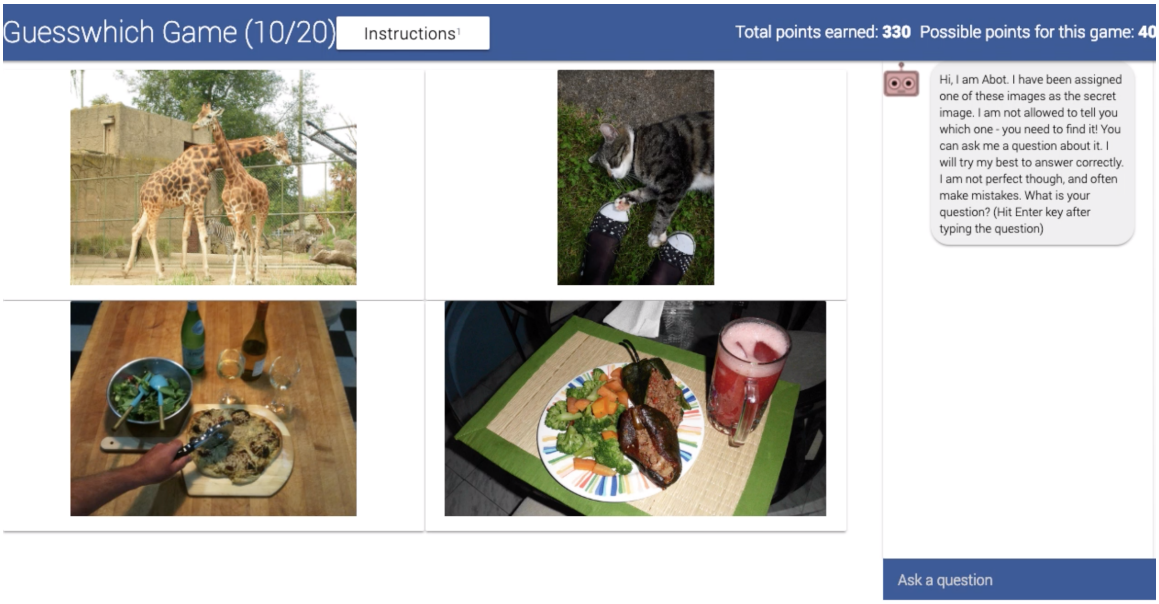
In this section, we describe details regarding the interactive game setup between humans and the AI. We first describe the relevant details regarding the GuessWhich game that we introduced in [41], and then describe our additions to this interface to evaluate the role of explanations in the context of the GuessWhich task.

### 7.1.1 Gameplay

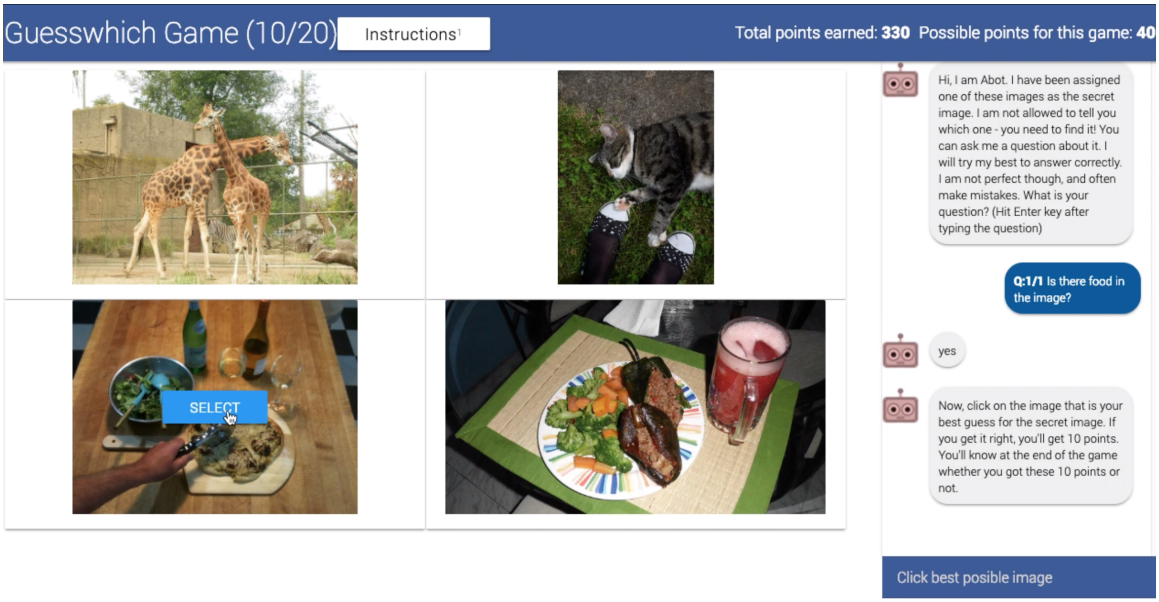
The AI is assigned a secret image from a pool of images sampled from the COCO dataset [125]. The identity of the secret image is unknown to the human subject – identifying this image from the pool is the goal of the task. Further details regarding pool construction are provided in Sec. 7.1.2. The subject attempts to guess the secret image from the pool by asking a question about the secret image. We provide this question as input to a VQA model along with the secret image. Details regarding the AI are described in Sec. 7.1.3. The output from the model, i.e., the answer string, is the AI's response to the subject. The subject then makes *one* guess of the secret image by clicking on the image that they believe to be the secret image.

Following the subject's guess of the secret image without explanations, we display explanations from the model on the interface. At this stage, we show the explanations by default but the subject has the option of toggling the explanations on or off via a button on the interface. In the case of Grad-CAM, the visual explanation is in the form of a heat-map that is overlaid on each of the images in the pool. The heat-map indicates the regions in the secret image that are most salient for the model's output. Fig. 29 and Fig. 30 show examples of a game without and with Grad-CAM explanations.

In the case of text explanations, the string of the text explanation is provided underneath each of the images in the pool, as shown in Fig. 31 and Fig. 32.




(a) Initial interface before subject's interaction.




(b) Subject clicks on an image based on the model's response to their question.

**Figure 31:** Screenshots of the game interface without explanations. The subject is first shown the pool of images. The subject asks a question to the model ('Is there food in the image?') based on the pool of images. The model responds to the subject's question ('yes'). The subject then selects an image (bottom left) as a guess for the secret image.

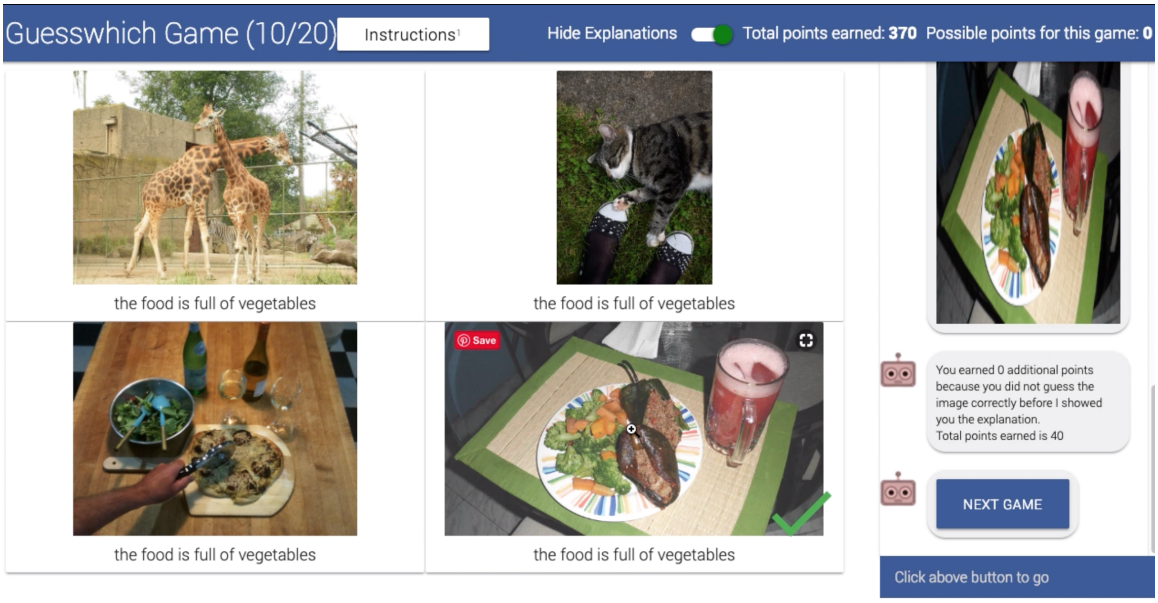



**Figure 32:** Screenshots of the game interface with text explanations. Once the subject selects an image as their guess, the rationale from the model in the form of a sentence is provided below each of the images. Based on the text explanations, i.e., the "reason why the model predicted the answer that it did", the subject then selects an image as a guess for the secret image.

The subject considers both the answer from the model, and the explanation regarding the answer, to guess the secret image. The user is allowed successive guesses (with feedback) until they guess the secret image correctly. We display a running score that the subject has achieved so far based on the accuracy of their guesses. We update this score at the end of each game. This serves as a technique to gamify the task to make it more interesting. It also serves as an incentive for the subjects to perform well since we provide them a bonus that is proportional to the game score.

Overall, each human subject plays 20 separate games (i.e., on 20 different pools) and attempts to guess the secret image before and after accounting for explanations from the model. Crucially, every task (of 20 games) across different experimental settings, is performed by a unique subject to prevent leakage of information regarding

the model across tasks.

### 7.1.2 Pool Selection

The pool of images are selected to ensure that the game is challenging, yet engaging for a human subject on a crowd-sourcing platform. In pilot studies, we found that the GuessWhich interface with a pool of 20 images resulted in a task that was too difficult. Following experimentation with varying pool sizes and rounds, we ultimately chose the number of images in the pool $N = 4$, and number of rounds of question-answering $R = 1$. The pool of images are sampled from the validation split of the COCO dataset [125] to avoid overlap with images that the VQA model was trained on. Images in their original aspect ratio are placed in random order in a rectangular grid, as shown in Fig. 29.

The pool is constructed by first sampling the secret (target) image from COCO. Based on the secret image, a hard-negative that is a 'neighbor' to the secret image is sampled. Then, two images are randomly sampled to complete the pool. We provide details regarding each of these below.

**Secret image**.

Each subject plays a total of 20 games. In an effort to create diverse games, we sample a diverse set of secret images since the construction of a pool of images is contingent on the secret image. First, we compute the average representation of all images belonging to a particular category in the COCO validation set. We consider this the 'canonical representation' for that category. The image whose representation is closest to the canonical representation, is considered a 'canonical image' for the category. The representation for an image is obtained by extracting the activations from the penultimate layer of VGG Net [210], a popular convolutional neural network.

Overall, we acquire 80 canonical images corresponding to each of the 80 COCO

categories. From among the canonical images, we sample 20 images uniformly randomly for use as secret images in our 20 games. Given the secret image for each game, we construct the rest of the pool as described below.

**Hard negatives**.

The GuessWhich game involves identifying the secret image from among distractor images. Thus, it is likely that the difficulty of the game depends on the number of images that are visually similar to the secret image. The more the number of images that are similar to the secret image, the harder it might be for the subject to eliminate the distractor images to correctly identify the secret image. In order to ensure that the game is sufficiently challenging, we sample one distractor image for each pool from among the neighbors of the secret image. Neighbors are identified based on small Euclidean distance in the image representation space, i.e., activations from the penultimate layer of a VGG-19 CNN [210].

**Random images**.

For each pool, we sample 2 images uniformly randomly from the COCO validation set. These images are distinct from the secret image and the hard negative (neighbor) image.

In summary, the pools are constructed by choosing 1 secret image, sampling 1 hard-negative, i.e., nearest neighbor image to the secret image, and sampling 2 images uniformly randomly from the COCO validation set.

### 7.1.3 AI

The AI that we implement in the experiments with Grad-CAM is a CNN-LSTM model [130] that is trained to perform the task of Visual Question Answering (VQA) [5]. The AI in the text explanations experiments is based on the Bottom-Up-Top-Down (BUTD) VQA model introduced by Anderson et al. [3] along with an explanation module, described by Wu and Mooney [243].

Both VQA models predict an answer, given an image and a question about the image. Details regarding each of the models are provided below.

**CNN-LSTM model**

The CNN-LSTM VQA model is a two-stream architecture that encodes the two modalities of the input – i.e., the image and the text question. The representations from each of these modalities is then fused together. The output from the model is based on the fused image-question representation. The image is encoded by a convolutional neural network (CNN), specifically, a VGG-19 Net [210], and the text question is encoded by a sequence model, specifically an LSTM [94] which is a type of recurrent neural network (RNN). The image representation from the encoder is the set of activations from the penultimate layer of a VGG-19 Net. The question representation consists of the activations from the hidden state of the LSTM at the end of last time-step of the sequence encoding. The question and image representations are transformed to the same size and fused using a point-wise multiplication.

The CNN-LSTM model outputs a score over N categories (answers), conditioned on the fused multi-modal representation. The output categories are the 1000 most frequent answers in the training set. The answer with the highest score is provided to the subject as the response of the AI.

The CNN-LSTM model achieves an accuracy of 58.16% on the test-standard of the VQA 1.0 dataset. Selvaraju et al. [203] implement Grad-CAM on the CNN-LSTM model and find that the Grad-CAM heat-maps from this model are weakly correlated with human attention maps (based on the rank correlation metric and human attention maps obtained from [53]). We use this model due to its simplicity and the interpretability of the heat-maps. We also tried visualizing Grad-CAM explanations from the Hierarchical Co-Attention model (a popular VQA model introduced by Lu et al. [131]), but from visual inspection, found the heat-maps to be less interpretable.

**Bottom-Up-Top-Down model**

In the experiments with text explanations, we employ the VQA model architecture described in [243] which is similar to BUTD [3]. In the BUTD model, a set of meaningful image regions are first proposed as candidates, and the visual features from each of these regions is then weighted using an attention mechanism. While in [3], object bounding boxes from Faster R-CNN [190] are considered the candidate regions, the architecture in our experiments uses segmented image regions from an image segmentation model [96]. The question representation is obtained by encoding the question via a Gated Recurrent Unit [47]. The question representation is used as context for attending over the visual features from the segmentation model. The weighted combination of the attended visual features is fused with the question representation using point-wise multiplication.

Conditioned on the fused multi-modal representation, the model outputs a score over $N$ categories (answers). There are ($N$ =) 3127 categories that includes every correct answer that appeared more than 8 times in the training set. The answer with the highest score is provided to the subject as the response of the AI.

We refer readers to Wu et al. [243] for further details regarding the VQA model and explanations module.

### 7.1.4 Grad-CAM

We experiment with the Grad-CAM explanation modality, which is a visual explanation that is faithful to the model's prediction. Recall from Sec. 6 that Grad-CAM is a saliency based approach that assigns weights to the regions in the image that contribute positively to any given prediction. An image is passed through a series of convolutions, and represented as a grid of 14x14 spatial regions. Given a particular output, Grad-CAM allows us to compute the support – in terms of spatial regions of the image – that contributes to the given output. These contributions are visualized

as a grayscale heat-map, which can be back-projected into the size of the original image. For further details regarding Grad-CAM, see [203].

In our experiments, we compute a Grad-CAM heat-map that explains the model's most confident prediction (i.e., the answer being displayed to the subject) on the secret image, given the subject's question. When displaying Grad-CAM explanations, we overlay the heat-maps (described earlier) over each image in the pool. Larger contribution weights correspond to lighter shade in the heat-map, and smaller contributions correspond to darker shades. Thus, when the Grad-CAM maps are overlaid on the original image, the regions of relatively high contribution are more visible than the regions with smaller contributions.

### 7.1.5 Text explanations

A recent interesting direction of research in explainable AI attempts to provide rationales regarding a model's prediction, in the form of a sentence [243]. The text rationale is generated by an 'explanation module' which is conditioned on the image, the VQA model's attention over the segmented objects, the input question to the VQA model, and the answer predicted by the VQA model. The explanation module is trained using human-written rationales that attempt to explain the VQA model's prediction. To determine whether a rationale is consistent with the VQA model's prediction, the visual features that contribute most to the rationale are identified via a sensitivity analysis. The rationales for which the highly contributing visual features also have high attention weights from the VQA model, are deemed to be consistent with the VQA model. The explanation module is trained only using the consistent human-written rationales. For further details, please see [243].

### 7.1.6 Human players

We recruit human subjects from Amazon Mechanical Turk (AMT)[1]. To ensure that we recruit subjects who read instructions carefully while performing tasks on AMT, we constrain our subjects to those who are that are located in the US, have completed at least 5000 HITs (tasks) on AMT, and have a task approval rating of above 98%. Each subject performs only a single task, consisting of 20 games to prevent the familiarity from these games from 'leaking over' to other tasks, leading to potentially inflated game scores. In order to motivate the subjects to perform well on the GuessWhich games, we provide subjects with a bonus reward that is proportional to their game score. The bonus is in addition to the fixed reward amount that the subjects are paid for performing the collaborative goal-driven task.

### 7.1.7 Infrastructure

The games involve setting up a live interaction between a human subject and a VQA model running on a GPU. The model's response (answer) is based on the human subject's question. Furthermore, we also generate the explanation for the model's answer, i.e., the Grad-CAM heat-map. To do this, we follow the setup described in detail by Chattopadhyay et al. [41].

## *7.2 Experiments*

In this section, we compare the overall performance of the human-AI team before and after explanations become available. Recall that the subjects playing the game attempt to guess the secret image in two stages – first, *only* based on the model's answer to the subject's question, and second, after the model's explanation regarding the answer is also provided. We compare the count of correct guesses of the secret image – given only the model's answer to their question, and given both the answer

[1] `https://www.mturk.com/`

and explanation. We report the fraction (in %) of correct guesses by all human subjects to the total number of games played.

### 7.2.1 Irrelevant visual explanation

To benchmark the performance of the human-AI teams at the task, we implement a baseline (control) experiment. For this baseline experiment, the Grad-CAM explanations are randomly generated via the following procedure – we sample a random question, and sample a random image from the COCO validation set. We provide this question and image as input to the VQA model. We then compute the Grad-CAM map corresponding to the most confident answer from the model for this random question about the image. We overlay this unrelated 'explanation' heat-map over the images in the pool, instead of the actual explanation.

Overall, 31 subjects play a total of 20 games each. I.e., we have a total of $31 * 20 = 620$ data points over which we compute average accuracy. We observe in Table 5 ('Irrel. Grad-CAM.') that the accuracy of the human subject in guessing the secret image correctly is exactly the same both before and after they take the explanations into account. We do observe that subjects do indeed change their mind on individual instances – i.e., on many occasions they do change their guess of the secret image after considering the (unrelated) explanation. Coincidentally, the total number of times the human subjects make a correct guess before explanations happens to be equal to the total number of correct guesses post-explanation.

We compute the improvement in performance after the explanation becomes available. Across subjects, the mean improvement acc. $+/-1.96*$std. error $= 0.000 + / - 3.060$. I.e., although the performance increases or decreases for each individual subject, the average difference in performance is 0.

In summary, as we would expect, random (unrelated) Grad-CAM explanations do not appear to improve performance of the human-AI team at the task. The

**Table 5:** Performance of the human-AI team at the goal-driven task before and after the human subjects gain access to explanations. Performance is measured in terms of mean accuracy (fraction of correct guesses to the total games played, in %) across games played all human subjects. The error terms are the 95% confidence intervals around the mean (1.96∗std. error). The experimental settings with irrelevant visual explanations (Irrel. Grad-CAM), relevant visual explanations (Rel. Grad-CAM), and relevant text explanations (Rel. text exp.) are described in Sec. 7.2.1, Sec. 7.2.2 and Sec. 7.2.3 respectively.

| | **Irrrel. Grad-CAM** | **Rel. Grad-CAM** | **Rel. text exp.** |
|---|---|---|---|
| Before exp. | 32.581 +/- 3.764 | 29.844 +/- 3.617 | 33.065 +/- 3.779 |
| After exp. | 32.581 +/- 3.764 | 34.219 +/- 3.751 | 47.742 +/- 4.012 |

performance does not significantly decrease either. I.e., it appears that the unrelated Grad-CAM map does not mislead the human subjects into choosing the wrong image on significantly many occasions.

### 7.2.2 Relevant visual explanation

In this experiment, we show subjects a relevant explanation – the Grad-CAM heat-map is computed for the subject's question and the model's most confident prediction (answer) for the secret image. In other words, the explanation highlights regions in the secret image that contribute to the model's answer to the subject's question. Overall, 32 subjects play a total of 20 games each.

The performance of the human-AI team in this experimental setting is presented in Table 5 ('Rel. Grad-CAM.'). First, we observe that there exists a difference between the performance of the human subjects across our actual experiment described earlier, and this baseline version *before* the explanation is shown. Since the VQA model is identical in both cases, there exists an inter-trial variance due to the variability in human subjects.

Second, the mean accuracy after the subject considers explanations, is higher than the mean accuracy before explanations are available to the subject. However,

there is also significant variance in performance across subjects, as indicated by the 95% confidence intervals around the mean. Another way to view this difference in performance is by directly computing the improvement in performance after the explanations becoming available. I.e., for each subject, we compute: avg. accuracy (across 20 games) after explanations - avg. accuracy before explanations. The mean improvement in accuracy across subjects $+/-1.96*$std. error $= 4.375 +/- 3.664$. Thus, we observe a slight but significant improvement in the subjects' ability to guess the secret image after they gain access to the relevant Grad-CAM explanations.

### 7.2.3 Relevant text explanation

In this experiment, subjects are provided the relevant text explanation – i.e., a rationale that explains the VQA model's predicted answer on the secret image, for the question provided by the human subject. Overall, 31 subjects play a total of 20 games each.

In Table 5 ('Rel. text exp.'), we first observe that the accuracy of the human-AI team before explanations is close to the performance of the CNN-LSTM model from the experiments with visual explanations. This is interesting, given the differences in architecture, and the datasets on which the two models were trained on. The CNN-LSTM model was trained on VQA v1, which is only a subset of the VQA v2 dataset that the BUTD model was trained on. In terms of performance, the BUTD model achieves a VQA accuracy of 70.34 on the VQA v2 test set, compared to the CNN-LSTM's 58.16 on the VQA v1 set (despite it being typically, an easier task). Thus, the differences in absolute performance of the VQA models do not seem to translate very well to performance in the collaborative human-AI task of GuessWhich.

Second, we note that the performance of the team is significantly higher after the subjects are provided with the text explanations. Specifically, we compute the improvement across subjects – the mean improvement acc. $+/-1.96*$std. error

$= 14.677 + / - 5.473$. Thus, text explanations seem to provide the human subject with information that helps them better identify the secret image. Indeed, comparison with Grad-CAM explanations suggests that the information provided by the text explanations is much more useful to the human subject compared to the visual explanations.

While the exact nature of the information that helps the human subject is an interesting open question, we present some analyses of the text explanations in Sec. 7.3, and identify a few open questions for future research in Sec. 7.4.

### 7.2.4 Crowd-sourcing details

As we mentioned earlier, we make use of the Amazon Mechanical Turk (AMT) crowdsourcing platform to recruit human subjects for our experiments. For each of the tasks, we paid the workers a fixed amount of $1.00, and a variable amount of bonus that is proportional to their performance in the task. The maximum possible bonus is $0.25. The median time of completion of one task (consisting 20 games) with relevant visual explanations is 14:54 min.

## *7.3 Analysis*

In this section, we present the accuracy of the VQA model in the GuessWhich game, and consider a few aspects of how the predicted answer may influence the subject's guess. We then study the extent to which certain characteristics of Grad-CAM heat maps such as intensity and spread are correlated with performance in the GuessWhich task.

### 7.3.1 Accuracy of VQA model

The performance of the human-AI team in the GuessWhich task (before explanations) depends on the pool of images, the question asked by the subject, the answer provided by the model, and the subject's ability to make the right guess. In this section, we

analyze the performance of the VQA model on the questions asked by the human subjects while playing the GuessWhich game.

We sample of 15 subjects randomly, from the set of 31 subjects who played the GuessWhich game with the Grad-CAM explanations. We manually verify if the answer by the CNN-LSTM VQA model for the subject's question on the secret image is correct or wrong. Overall, we annotate the accuracy of the VQA model for $15*20 = 300$ games. We find that the mean accuracy of the model is $53.33\% +/- 4.46\%$ (error is the 1.96∗standard error of the mean that corresponds to a 95% CI). Note that the human subject guesses the secret image correctly before explanations, approximately 29-32% of the time (Table 5).

While it is important for the VQA model to accurately answer the subject's question for the secret image, accuracy alone does not reflect the various aspects that are involved in making an accurate guess in the GuessWhich game. A few considerations are presented below:

1. Even when the model provides an answer that is technically incorrect, it could be informative. For instance, consider the question, "What is in the image?". For an image that contains a sandwich, the model answers "hot dog", which although incorrect, does provide a hint regarding the secret image. In another game, we observed that for the question, "What is in the image?" asked for the secret image containing a boat in water under a cloudy sky, the model responded with "overcast". Although the answer was incorrect in both these cases, the answers seemed helpful to guess the secret image.

2. We observe that a common strategy among human subjects is to form a hypothesis regarding the secret image, then confirm this hypothesis by asking yes/no questions. While the VQA model performs quite well on some questions that elicit information about the image, like 'What is in the image?', the model is not very accurate on yes/no questions. Despite the fact that subjects get feedback

**Table 6:** Mean fraction (in %) of high intensity pixels ($\mu_{I>\tau}$) in Grad-CAM heat maps across all games in the dataset where subjects guess the secret image correctly/incorrectly before/after explanations. The error terms are the 95% confidence intervals around the mean (1.96∗std. error). The number of games belonging to each category are given in the parenthesis (each row sums to the total of 620 games). $\mu_{I>\tau}$ and other details are described in Sec. 7.3.2.

| | **Correct** | **Incorrect** |
|---|---|---|
| Before exp. | 4.729 +/- 0.451 (185) | 5.087 +/- 0.306 (435) |
| After exp. | 4.959 +/- 0.420 (209) | 4.991 +/- 0.317 (411) |

at the end of every round that exposes the model's inaccuracy, most subjects continue their strategy of asking yes/no questions for all the 20 games that they play. This suggests that subjects do not often alter their question-asking strategy based on the model's weaknesses or inaccuracy. A possible reason for such a strategy might be due to their bias from the real-life 20-questions game where people are typically only allowed to ask yes/no questions.

3. An answer, although technically correct for the secret image, might not be a useful in the context of the pool. This is especially the case for yes/no questions that are not perfectly discriminative. For instance, consider the question, "Are there books in the image?" for a pool containing two images – one where the books are salient (distractor image), and another where the books are in the background (secret image). The model, which only has access to the secret image, answers "yes". This answer, however, can be misleading to the subject who might expect the model to respond in a pragmatic manner.

### 7.3.2 Grad-CAM intensity

Recall that in our experiments, Grad-CAM highlights locations in the secret image that contribute to the AI's answer. In this section, we perform some analyses in an attempt to identify characteristics of the Grad-CAM maps that are correlated with

**Table 7:** Mean fraction (in %) of high intensity pixels ($\mu_{I>\tau}$) in Grad-CAM heat maps for each game played by a subject. The error terms are the 95% confidence intervals around the mean (1.96∗std. error). The number of games belonging to each category are given in the third column (total number of games = 620). $\mu_{I>\tau}$ and other details are described in Sec. 7.3.2.

| **Accuracy** | $\boldsymbol{\mu_{I>\tau}}$ | **# games** |
|---|---|---|
| Incorrect before, Incorrect after exp. | 4.982 +/- 0.339 | 373 |
| Incorrect before, Correct after exp. | 5.718 +/- 0.651 | 62 |
| Correct before, Incorrect after exp. | 5.077 +/- 0.854 | 38 |
| Correct before, Correct after exp. | 4.640 +/- 0.522 | 147 |

performance at the GuessWhich game. We first examine the intensity of the Grad-CAM heat maps. The intensity at a particular location of a Grad-CAM heat map corresponds to the extent to which it contributes to the prediction of the answer. Locations with higher intensity contribute more, and lower intensity contribute to a lesser extent. We study trends of the number of high intensity locations (i.e., number of locations that contribute to the prediction to a significant extent), with performance at the GuessWhich task.

We first compute the approximate area in a Grad-CAM heat map that is covered by high intensity locations. We deem a pixel (location) to be of high intensity if the pixel intensity at that location exceeds a threshold $\tau$. $\tau = 118$ corresponds to the 95th percentile of pixel intensities across all Grad-CAM maps in our experiments. We find that for each Grad-CAM map, the fraction of pixels that is high intensity ($\mu_{I>\tau}$) $= 4.98\% +/- 0.25\%$ (mean $+/-$ 1.96∗ std. error).

We computed $\mu_{I>\tau}$ for each of the following settings – when the subject guessed the secret image correctly before explanations, wrongly before explanations, correctly after explanations and wrongly after explanations. We observe in Table 6 that the fraction is similar to the overall $\mu_{I>\tau}$ across all Grad-CAM heat maps (around 5%).

We also computed trends comparing finer-grained gameplay. Specifically, we consider the trend in *each game* and compute $\mu_{I>\tau}$ for each of the four possibilities shown in Table 7 – when a subject guesses the secret image correctly both before and after they see explanations (in the same game), when they guess it wrongly before explanations but correctly after, and so on. Interestingly, we observe that compared to all other games, $\mu_{I>\tau}$ is higher for the set of games where subjects guessed the secret image incorrectly before explanations, but correctly after (row 2). This difference is significant when compared with the games where the human subjects' guesses were either correct or incorrect, both before and after explanations (rows 1 and 4). This suggests that in cases where the subject's guess before explanations is incorrect, Grad-CAM maps with high $\mu_{I>\tau}$ are correlated with a correct guess after explanations.

A point to note regarding the above analysis and trends are that they depend on the choice of $\tau$. We expect that a very small $\tau$ (most pixels are considered high intensity) or a very large $\tau$ (very few pixels considered high intensity) would not be meaningful to study the correlation between the fraction of high intensity locations and performance. We choose a reasonable intermediate value, i.e., the 95th percentile of intensities.

### 7.3.3 Grad-CAM spread

In a similar vein to our previous analysis, we investigate if there exists a correlation between the 'spread' of the Grad-CAM heat-maps and performance of the human-AI team. To compute the spread of a Grad-CAM heat map, we first binarize the image using the threshold $\tau$ that we described in the section above – pixels with intensity above $\tau$ are set to 1 and the rest to 0. We then down-sample the image via a max-pooling operation – we use a max-pool kernel of size $(28, 28)$ to arrive at an $8 * 8$ grid of pixels containing the maximum intensities from their respective receptive fields of the original $224 * 224$ image. Finally, we find the number of connected components in

**Table 8:** Mean spread of high intensity pixels ($S_{>\tau}$) in Grad-CAM heat maps for each game played by a subject. The error terms are the 95% confidence intervals around the mean (1.96∗std. error). The number of games belonging to each category are given in the third column (total number of games = 620). $S_{>\tau}$ and other details are described in Sec. 7.3.3.

| **Accuracy** | $\mu_{I>\tau}$ | **# games** |
|---|---|---|
| Incorrect before, Incorrect after exp. | 2.115 +/- 0.115 | 373 |
| Incorrect before, Correct after exp. | 2.500 +/- 0.256 | 62 |
| Correct before, Incorrect after exp. | 2.289 +/- 0.364 | 38 |
| Correct before, Correct after exp. | 1.966 +/- 0.165 | 147 |

this binarized, downsampled image, and call it the spread score ($S_{>\tau}$). The intuition is that a Grad-CAM heat map with higher spread of high intensity pixels results in larger numbers of connected components. A heat map with a smaller spread of high intensity pixels results in smaller number of connected components.

We observe that the mean spread score $S_{>\tau}$ is similar (around 2.1) for the settings when the subject guessed the secret image correctly before explanations, wrongly before explanations, correctly after explanations and wrongly after explanations.

Similar to the intensity analysis presented above, we also compute the spread score of high intensity pixels for each game (see Table 8). We observe that the trends are similar to the intensity analysis, but less significant. Specifically, $S_{>\tau}$ for the set of games where the subject's guess was incorrect before explanations and correct after (row 2), is higher than all other sets of games. This is especially the case when compared with the games where the guesses are either both incorrect or correct both before and after explanations (rows 1, 4 respectively). However, in comparison with the games where the subject guessed correctly before but incorrectly after explanations, the difference is not as significant.

As we discussed earlier, we note that the trends that we present are likely sensitive to the threshold $\tau$, and the kernel size for the max-pooling operation.

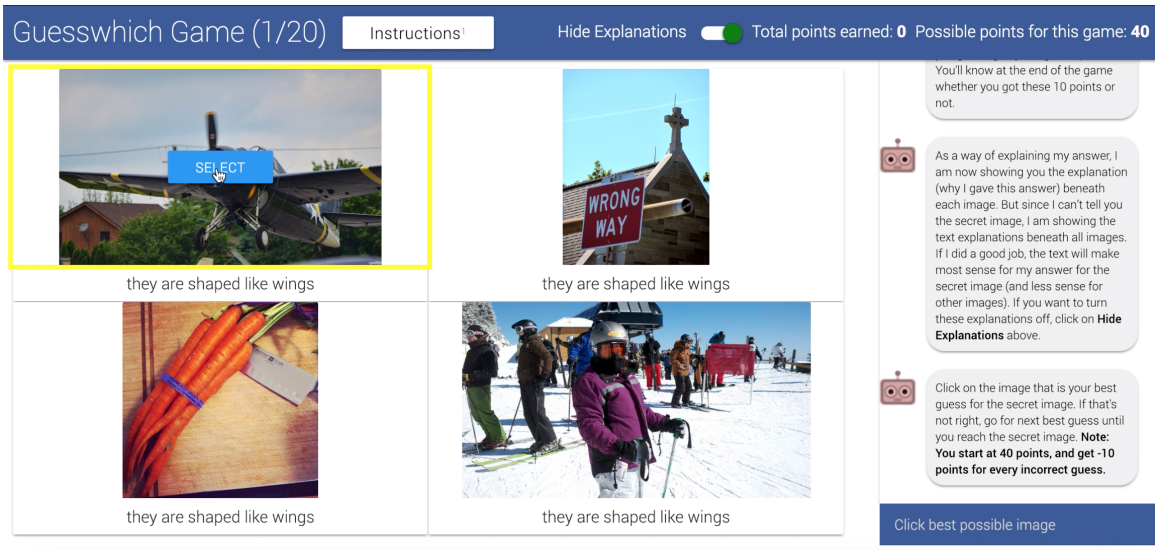



**Figure 33:** Screenshot of the game interface with a text explanation that is relevant to the secret image but not directly relevant to the question-answer. The text explanation provides the subject with 'extra' information regarding the secret image.

### 7.3.4 Text explanations provide 'extra' information

Recall from the results presented earlier in Sec. 7.2 that performance in the GuessWhich task improves with text explanations. Interestingly, we observe that at times, the information provided by the text explanations is not directly relevant to the question-answer from the model. Consider the game in Fig. 33. The subject asks the question, 'What objects are in the image?' for which the model's response is 'trees'. This answer alone might not be sufficiently discriminative for the human subject since trees are not the salient object in any of the images. Further, trees are present in the background of two images in the pool. Although the answer is ambiguous, the text explanation, 'they are shaped like wings' clearly provides the subject with a clue regarding the secret image – the top-left image containing an airplane. We note that while the answer was 'trees', the explanation is clearly referring to the airplane in the image. This raises an interesting and open question regarding measuring the consistency of a text explanation with the image, question, and answer.

## *7.4 Discussion*

In this section, we discuss a few considerations, nuances and caveats regarding the GuessWhich task, experimentals and performance that we report. We also identify interesting open questions for future work.

### 7.4.1 Chance performance

Recall the process of pool construction presented earlier in Sec. 7.1.2. This strategy to construct the pool has an unintended consequence – on observing the pool of images, it is often possible to discern that a pool contains two images that are closely related (the secret image and hard negative often belong to the same class), and two other images that are apparently unrelated to all other images.

On observing the pool of images, in the event that the human subjects recognize this similarity, and further reason that the secret image is one among the two similar images, we would expect the performance to be $\geq 50\%$. In our experiments, we observe that the mean performance (before explanations) is around the range of 29% to 32%. Thus, on average, the human subjects in our experiments do not seem to accurately identify our strategy for pool construction. A point to note is that the performance that we report is measured after one round of question-answering, i.e., the user first observes the pool, then asks a question, and guesses the secret image based on the model's answer to the question.

To accurately estimate the extent to which human subjects are able to identify our strategy for pool construction, one could perform the following experiment: given only the pool, ask human subjects to guess the secret image. In our experiments, we did not perform this experiment for two reasons: a) we are interested in performance of the human-AI team based on interactions between the human subject and the VQA model. As a result, we refrained from encouraging human subjects from analyzing the pool by itself, and b) our primary goal in these experiments is to evaluate the utility

of explanations in improving performance of the human-AI team. When evaluating the *improvement* in performance with the information provided by explanations, the potential added benefit of recognizing the pool construction mechanism would likely not change this difference. Thus, estimating the extent to which human subjects identify our strategy for pool construction is orthogonal to our primary goal in these experiments.

### 7.4.2 Sensitivity to choice of distractors

The main motivation behind the mechanism for pool construction (described in Sec. 7.1.2) is to ensure that the GuessWhich games that are of moderate difficulty. In previous work [41], and through pilot experiments, we found that a pool with a large number of images, especially, a large number of hard-negatives makes the GuessWhich game too difficult for human subjects. The reasonable rate of success of the human-AI teams (reported in Sec. 7.2), suggests that the games that are played with the constructed pools are neither too easy nor too difficult (given the VQA model that the humans interact with). Based on our experience from playing the GuessWhich game with explanations, we formed a few hypotheses regarding the gameplay and performance of the human-AI team, which we present below.

Consider the situation before explanations, when the model's answer for the subject's question on the secret image is correct. We hypothesize that the human subject can narrow down the plausible candidate images to the secret image and the hard-negative. Since the hard-negative is relatively close (in representation space) to the secret image, we expect that an answer to a question on the secret image, might also be applicable to the hard-negative image in many instances. This also holds true when the question on the secret image is answered incorrectly by the model. In this case, it is plausible that the human subjects narrow down the candidates to the two random images.

We further hypothesize that in the case where the the human subject narrows down the plausible candidates to two images – one of which is the secret image – they can effectively utilize the explanation from the model to accurately identify the secret image. Recall that the explanation available to the subject is with respect to the answer predicted by the model for the question asked on the secret image. Thus, in cases where the explanation is consistent with the predicted answer, question, and the secret image, we expect human subjects to accurately identify the secret image. Specifically, human subjects will likely be able to guess the secret image correctly when the explanation is more consistent with the predicted answer for the secret image than with other plausible candidate images.

In cases where the explanation is equally (or more) consistent with a candidate (distractor) image compared to the secret image, we do not expect the human subject to accurately identify the secret image.

Thus, we note that the performance of the human-AI team in this task is likely also sensitive to the pool of images, i.e., the choice of secret image and the distractors (hard-negative and random images).

### 7.4.3 Saliency vs. 'additional-information' explanations

Grad-CAM, the interpretable visual explanation that we utilize in our work is a saliency-based approach which highlights the relevant regions in the input that contributes to the model's most confident prediction. On the other hand, the text explanation provides a rationale for the predicted answer. This rationale is generated by an explanation module that is trained using the set of human-provided rationales that are relevant to the model's prediction for each given training sample. A text explanation is conditioned on (and hence intended to be consistent with) the specific objects in the image that the VQA model attended to. Note though, that unlike Grad-CAM, these regions are not necessarily the most salient for the model's prediction.

The text explanations are a fundamentally different type of explanation for the model's current prediction, compared to Grad-CAM. Apart from the difference in the modality in which the explanation in conveyed (text vs. visual), there exists another difference – while Grad-CAM provides a heat-map over the input which highlights the salient regions, the text explanations conveys a plausible rationale for arriving at the final decision. To a human subject, this is an explanation that attempts to provide a conceptual, textual rationale to the question of 'why' this prediction. In contrast, Grad-CAM does not provide a rationale for the model's prediction; it highlights 'what' region in the input the model focused on. It would be interesting to identify the specific instances, and applications where each of these types of explanations are useful, and for what reasons. Characterizing the ways in which humans consider and integrate information from these two different types of interpretable explanations, is an interesting open question.

### 7.4.4 Text explanations – sentence vs. bag of words

In this work, we experiment with text explanations that are full sentence rationales. While we find that these explanations do indeed improve performance of the human-AI team in the GuessWhich task, it is unclear if a full sentence rationale is necessary. I.e., is it possible that a subset of the full sentence, e.g., word(s), phrases can also serve as equally good (or better) explanations? If yes, then what are the characteristics of these words/phrases?

This is a useful question that has consequences in training the explanation module. For instance, in the event that full sentence explanations are just as (or less) useful compared to explanations in the form of a single word or phrase, this could result in a drastic reduction in the amount of rationale training data required for an explanation module. Further, it is possible that the explanation module that is trained on the shorter explanations learns better (higher accuracy on training data), and generalizes

better (higher accuracy on validation data).

In the event that full-sentence text explanations (consisting of similar words as a word/phrase based explanation) do indeed result in better performance of the human-AI team, it would be interesting to evaluate the specific reason for this improvement. One hypothesis is that full-text explanations from a model might be perceived by the human subject as more sophisticated, which might lead to better trust from the subject.

Another important line of inquiry could indeed question the premise that the explanation module needs to be learned from human-provided rationales. Specifically, is it possible to derive the text (word, phrase, or sentence) explanation from the final answer? For instance, consider the question, answer, and explanation in Fig. 32. The explanation of 'vegetables' is a hypernym of 'food', that is relevant to the secret image. Is it the case that the explanations from the model can simply be synonyms, and hypernyms that are related to the concept being referred to in the question and answer? How useful might these explanations prove to be in a downstream task like GuessWhich? These are interesting open questions that would likely improve our understanding of the utility of different types of explanations.

### 7.4.5 Counterfactual explanations

Another interesting direction of inquiry is regarding the utility of counterfactuals in the GuessWhich game. Specifically, can counterfactual explanations (i.e., explanations for an alternate answer) provide the human subject with additional information regarding the secret image?

In future work, we could evaluate this via the following experimental setup – given a pool of images and the human subject's question, first present the subject with the response and explanation for the secret image. Then, present the subject with a counterfactual, i.e., an alternate answer and the explanation corresponding to that

alternate answer. While the first answer and explanation provide some information to the subject regarding the secret image, the counterfactual answer and corresponding explanation might provide additional information regarding the secret image. Concretely, considering Grad-CAM explanations, the counterfactual explanation would provide a heat-map highlighting the regions in the secret image that support the alternate answer to the question. Depending on the relevance of the alternate answer to the secret image (and the explanation), this additional information may be useful to the human subject.

Alternately, consider the pool in Fig. 32 as an example. The subject's question 'Is there food in the image?' results in the answer 'yes', and the explanation is about the food – 'the food is full of vegetables'. Consider a counterfactual answer to this question – 'no'. A good explanation for such an answer, would likely describe objects in the image that are not food, e.g., drinks, table, etc. The additional information regarding these other objects in the image might help the subject better guess the secret image.

An interesting question that we left unanswered in the above discussion is, what is a 'useful' alternate answer? The alternate answer could be the second most confident answer predicted by the VQA model. On the other hand, we could also elicit an alternate answer from the human subject. I.e., the human subject could 'ask' the model for a heat-map highlighting the regions in the secret image that support the human subject's answer for their question. The utility of different versions of alternate answers is an open empirical question for future work.

Providing counterfactual visual explanations (i.e., using Grad-CAM) would involve a simple redesign of the current interface. However, generating counterfactual text explanations using our current framework might be a little more challenging, and constitutes an interesting research question.

## *7.5 Conclusion*

In this work, we evaluate the extent to which a visual explanation modality influences the performance of a human-AI team in a collaborative, goal-driven task. We propose and implement an image-guessing game as an instantiation of a goal-driven collaborative task. As an explanation, we use Grad-CAM, a saliency-based approach to visualize the regions in the image that contribute positively to the current prediction. We also experiment with text-explanations that provides a rationale regarding the model's answer. In experiments, we observe that performance at the goal-driven task improves after explanations are made available to the human subject, and that the improvement is larger for the text explanations than the visual explanations. Our work demonstrates the potential of using explanations to improve performance of a human-AI team performing a collaborative task.

Developing approaches that can make AI more interpretable, designing evaluation methods to quantify the utility of these approaches is an important and active area of research. In the context of the increasing interest in interpretable and explainable AI, we propose an additional method to measure the utility of an explanation – in terms of the improvement on a concrete goal-driven human-AI task. We suggest that goal-driven tasks such as the image-guessing task can be a good test bed to evaluate generic explanation modalities.

# CHAPTER VIII

# CONCLUSION

In this dissertation, I modeled and evaluated some aspects of inter-human interactions, with the goal of making human-AI interactions more natural, i.e., more similar to human-human interactions. With the increasing use of AI in applications that interface with humans, it is essential that the technology interacts with people in a way that is natural to them.

Inter-human interaction is rich, involving pragmatics, humor, sarcasm, empathy, sympathy, story telling, and leveraging a good theory of mind. In this dissertation, I focused on three broad areas that involved aspects of humor, story-telling, and theory of (AI's) mind in the domains of vision and language. Specifically, I

1. Built computational models for humor manifested in static images (described in Sec. 3), and contextual, multi-modal humor (described in Sec. 4).

2. Introduced a picture-sequencing task where a computational model learns the correct temporal order of events in a story (Sec. 5).

3. Evaluated different factors that influence the extent to which a lay person can predict the behavior of an AI (described in Sec. 6).

4. Evaluated the influence of an interpretable visual explanation that explains the AI's decision, on the overall performance of a human-AI team in a goal-driven, cooperative task (Sec. 7).

I believe that natural interactions that include aspects of humor, narrative, and predictability will likely improve the usability of technology for a lay person, and have the potential to increase the effectiveness of the collaborating human-AI team.

# APPENDICES

# APPENDIX A

# VISUAL HUMOR

In the following appendix we provide:

1. Inter-human agreement on funniness ratings in the Abstract Visual Humor (AVH) dataset.
2. Details of the model architecture used to learn object embeddings and visualizations of its embeddings.
3. A sample of objects from the abstract scenes vocabulary.
4. Examples of scenes from our datasets.
5. Analysis of occurrences of different object types in scenes from our datasets.
6. The user interfaces used to collect scenes for the AVH and Funny Object Replaced (FOR) datasets.

## *A.1 Inter-human Agreement*

In this section, we describe our experiment to determine inter-human agreement in funniness ratings of scenes. The Abstract Visual Humor (AVH) dataset contains 3,028 funny scenes and 3,372 unfunny scenes that were created by Amazon Mechanical Turk (AMT) workers. The funniness of each scene in the dataset is rated by 10 different workers on a scale of 1-5. We define the *funniness score* of a scene, as the average of all ratings for a scene. In this section, we investigate the extent to which people agree regarding the funniness of a scene.

Perception of an image differs from one person to another. Moran et al. [159] treat humor appreciation by people as a personality characteristic. We investigate to

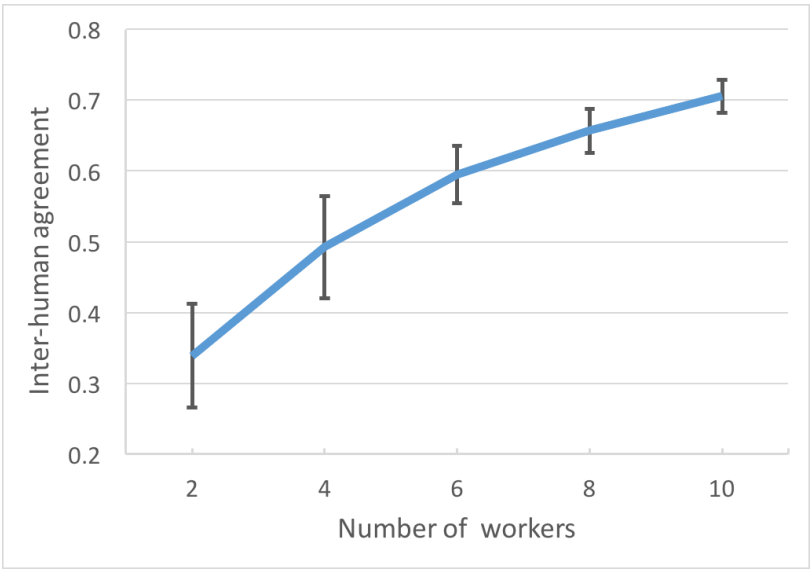



**Figure 34:** Inter-human agreement (y-axis) as we collect funniness ratings from more workers (x-axis). We see can see that by 10 ratings, we are starting to saturate with high agreement, indicating that 10 ratings is sufficient for a reliable *funniness score*.

what extent people agree how funny each scene in our dataset is. We split the votes we received for each scene into two groups, keeping each individual worker's ratings in the same group to the extent possible. We compute the *funniness score* of each scene across workers in each group. We compute Pearson's correlation between the two groups. Fig. 34 shows a plot of Pearson's correlation (y-axis) *vs*. the number of workers (x-axis). We can see that inter-human agreement increases as we increase the number of workers in a group and that the trend is gradually saturating. This indicates that ratings from 10 workers is sufficient to compute a reliable *funniness score*.

We observed that the standard deviation among ratings from 10 different workers for funny scenes is 1.09, and for unfunny scenes is 0.73. I.e., people agree more on scenes that are clearly not funny than on ones that are funny, matching our intuition that humor is subjective, while the lack thereof is not.

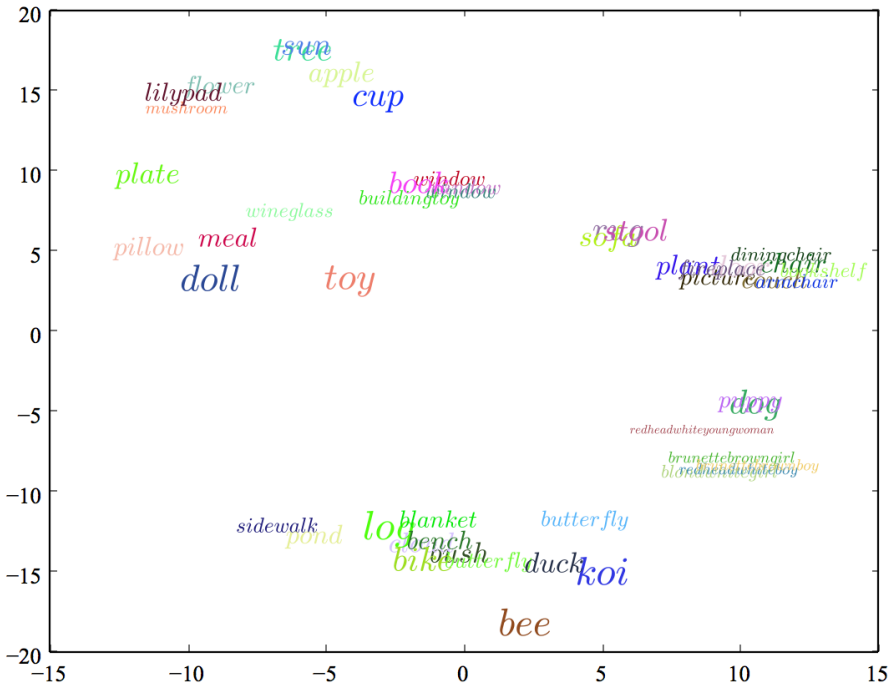



**Figure 35:** Visualization of 'normal' object embeddings of 75 most frequent objects in unfunny scenes. We see that closely placed objects have semantically similar meanings.

## *A.2 Object Embeddings*

In this section, we describe our model that learns embeddings for clipart objects and present visualizations of these embeddings. We learn distributed representations for each object category in the abstract scenes vocabulary using a word2vec-style continuous Bag-of-Words model [149]. During training, subsets of 6 objects are sampled from all of the objects present in a scene and the model tries to predict one of the objects, given the other 5. Each object is assigned a 150-d vector, which is randomly initialized. The vectors corresponding to the 5 context objects are projected to an embedding space via a single layer whose parameters are shared between the 5 objects. This (randomly initialized) layer consists of 150 hidden units without a non-linearity after it. The sum of these 5 object projections is used to compute a softmax over the 150 classes in the object vocabulary. Using the correct label (i.e., the object category of the 6th object), the cross-entropy loss is computed and backpropagated to learn all

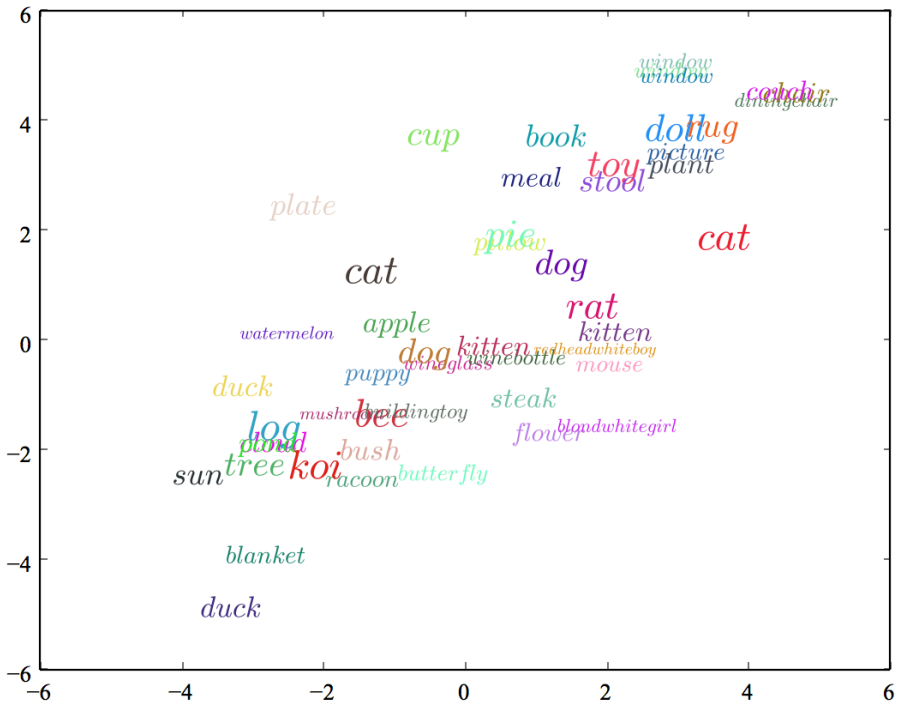



**Figure 36:** Visualization of 'humor' embeddings of 75 most frequent objects in funny scenes. We see that objects that are close in the 'humor' embedding space may be semantically very different.

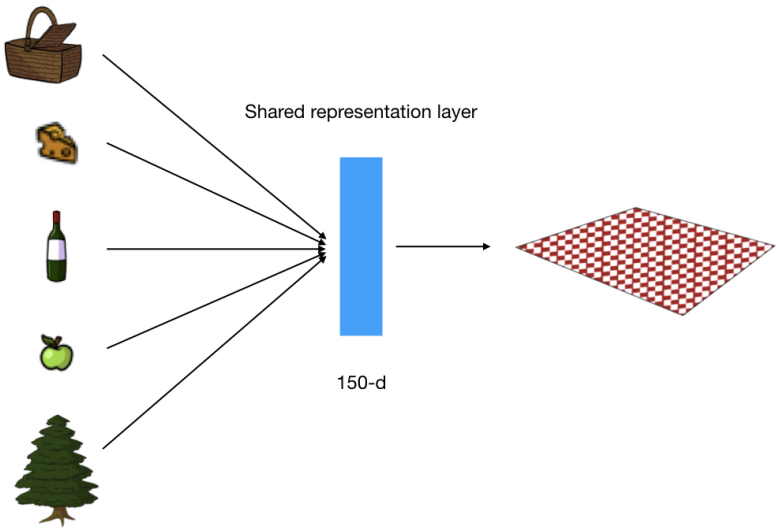



**Figure 37:** The continuous Bag-of-Words model projects 5 objects in the scene into a 150-d representation space. The 6th object in the scene is predicted, given the sum of the representations from these objects.

network parameters. The model is trained using Stochastic Gradient Descent with a base learning rate of 0.0001 and a momentum update of 0.9. The learning rate was reduced by a factor of two after each epoch. A diagram of the model can be seen in Fig. 37.

The context provided by the 5 objects ensures that the representations learnt reflect the relationships between objects. I.e., objects that are semantically related tend to have similar representations. We learn the 'normal' embeddings (i.e., the object embedding instance-level features from the main paper) from 11K scenes collected by Antol et al. [5]. As these scenes were not intended to be humorous, the relationships captured in the embeddings are the ones that occur naturally in the abstract scenes world.

Fig. 35 *(left)* is a t-SNE [57] visualization of the 'normal' embeddings for the 75 most frequent objects in unfunny scenes. In Fig. 35 *(right)*, we also visualize 'humor' embeddings, which were not used as features but provide us with insights. These are learnt from the 3,028 funny scenes in the AVH dataset.

We observe that the 'normal' embeddings encode a notion for which object categories occur in similar contexts. We also observe that closely placed objects in the 'normal' embedding space have semantically similar meanings. For instance, humans are clustered together around coordinates *(10, -7)*. Interestingly, `dog` and `puppy` (coordinates *(10, -5)*) are placed together and furniture like `chair`, `bookshelf`, `armchair`, etc. are placed together (coordinates *(10, 5)*). This follows from the distributional hypothesis, which states that words which occur in the similar contexts tend to have similar meanings [72, 90].

In contrast, in the 'humor' embeddings, visualized in Fig. 35 *(right)*, we see that objects that are close in the embedding space may be semantically very different. For instance, `dog` and `wine glass` are placed together at coordinates *(0, 0)*. These

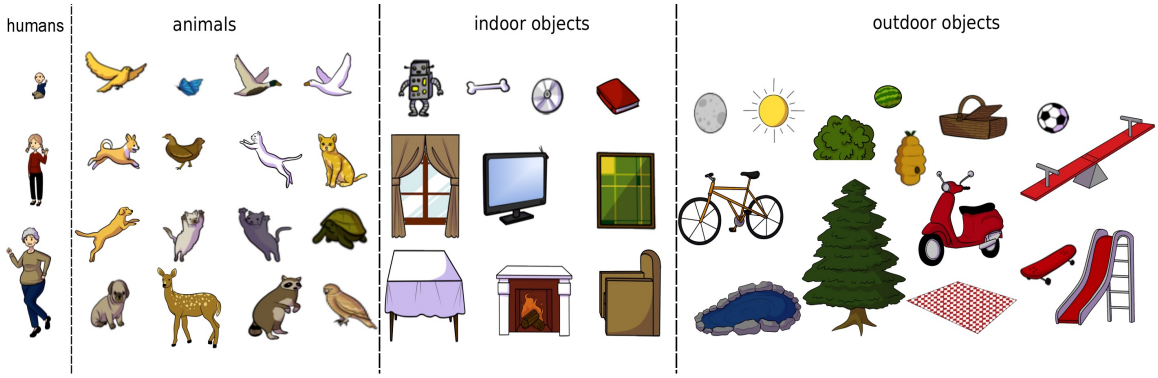



**Figure 38:** A subset of clipart objects from the abstract scenes vocabulary.

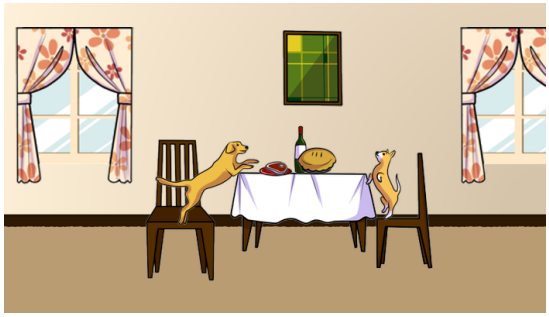
are placed far apart (at opposite ends) in the normal embedding. However, in the humor embedding, these two categories are extremely close to each other; even closer than semantically similar categories like two breeds of dogs. We hypothesize that this because our dataset contains funny scenes consisting of dogs with wine glasses, e.g., Fig. 39(b). It is interesting to note that 'background' objects that do not contribute to humor in a scene are also placed together. For example, `chair`, `couch`, and `window` are placed together in the humor embedding as well (coordinates *(4, 5)*).

The understanding of semantically similar object categories that can occur in a context, represented by the normal embeddings, can be interpreted as a person's mental model of the world. The humor embeddings capture deviations or incongruities from this normal view that might cause humor.

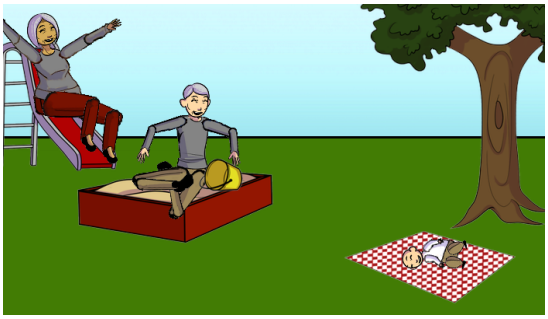
## *A.3 Abstract Scenes Vocabulary*

The abstract scenes interface developed by Antol et al. [5] consists of 20 'deformable' humans, 31 animals in different poses, and about 100 objects that can be found in indoor scenes (*e.g.*, couch, picture, doll, door, window, plant, fireplace) or outdoor scenes (*e.g.*, tree, pond, sun, clouds, bench, bike, campfire, grill, skateboard). In addition to the 8 different expressions available for humans, the ability to vary the

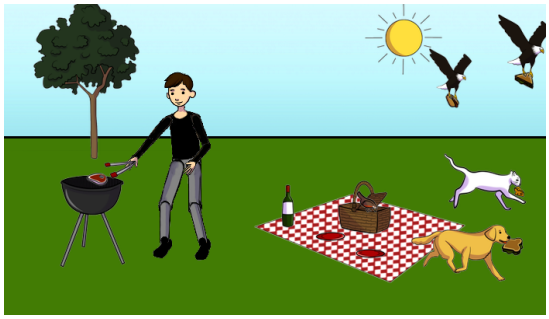
(a) 1.3



(b) 2.8



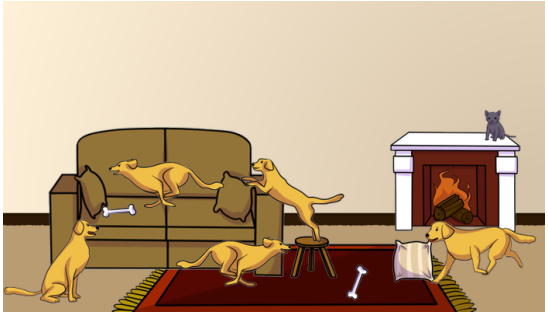
(c) 3.2



(d) 4.4

(e) 1.1

(f) 2.7



(g) 3.5in

(h) 4.1

**Figure 39:** Spectrum of scenes from our AVH dataset that are arranged in ascending order of *funniness score* (shown in the sub-caption)

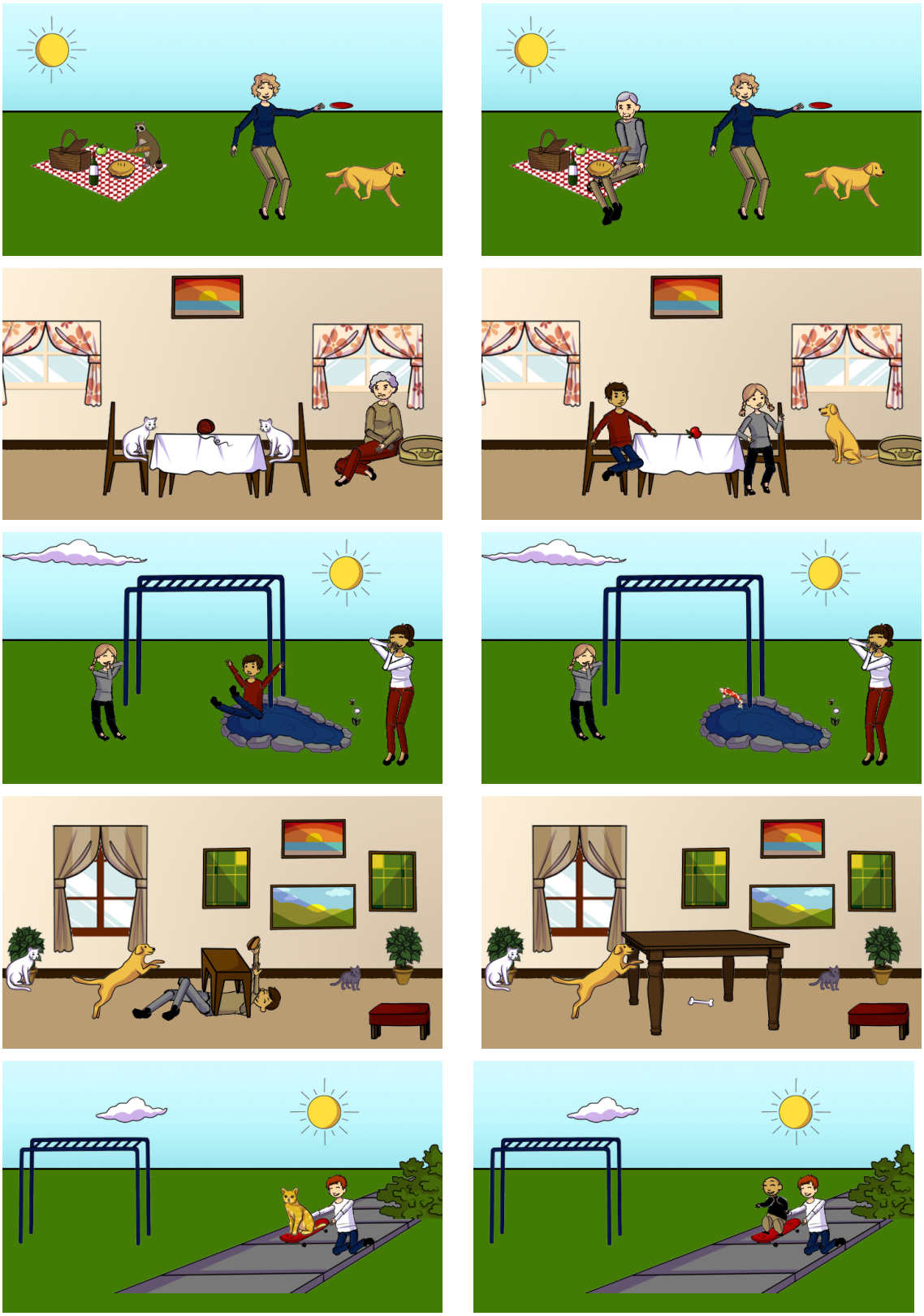


**Figure 40:** Some example originally funny scenes (*left*) and their object-replaced unfunny counterparts (*right*) from the FOR dataset.

pose of a human at a fine-grained level enables these abstract scenes to effectively capture the semantics of a scene. The large clipart vocabulary (of which only a fraction is shown to a worker during creation of a scene) ensures diversity in the scenes being depicted. A subset of objects from our Abstract Scenes vocabulary is shown in Fig. 38.

## *A.4 Example Scenes*

In this section, we present examples of scenes that were created using the abstract scenes interface. Fig. 39, depicts a spectrum of scenes from the AVH dataset in ascending order of *funniness score*. These scenes were created by AMT workers using the interface presented in Fig. 43.

Fig. 40 shows originally funny scenes (*left*) and their unfunny counterparts (*right*) from the FOR dataset. AMT workers created the counterparts by replacing as few objects in the originally funny scene such that the resulting scene is not funny anymore. A screenshot of the interface that was used to create the unfunny counterparts is shown in Fig. 44.

## *A.5 Object Type Occurrences*

In this section, we first analyze the occurrence of each object type in funny and unfunny scenes. We then analyze the most commonly cooccurring object types in funny scenes as compared to unfunny scenes.

**Distribution of Object Types.** We analyze the distribution of object types in funny and unfunny scenes across all scenes in our dataset. We compute the frequency of appearance of each object type in funny and unfunny scenes. We use this to compute the probability of a scene being funny, given that an object is present in the scene, which is shown in *blue* in Fig. 42. Since we have more unfunny scenes than funny scenes, we use normalized counts.

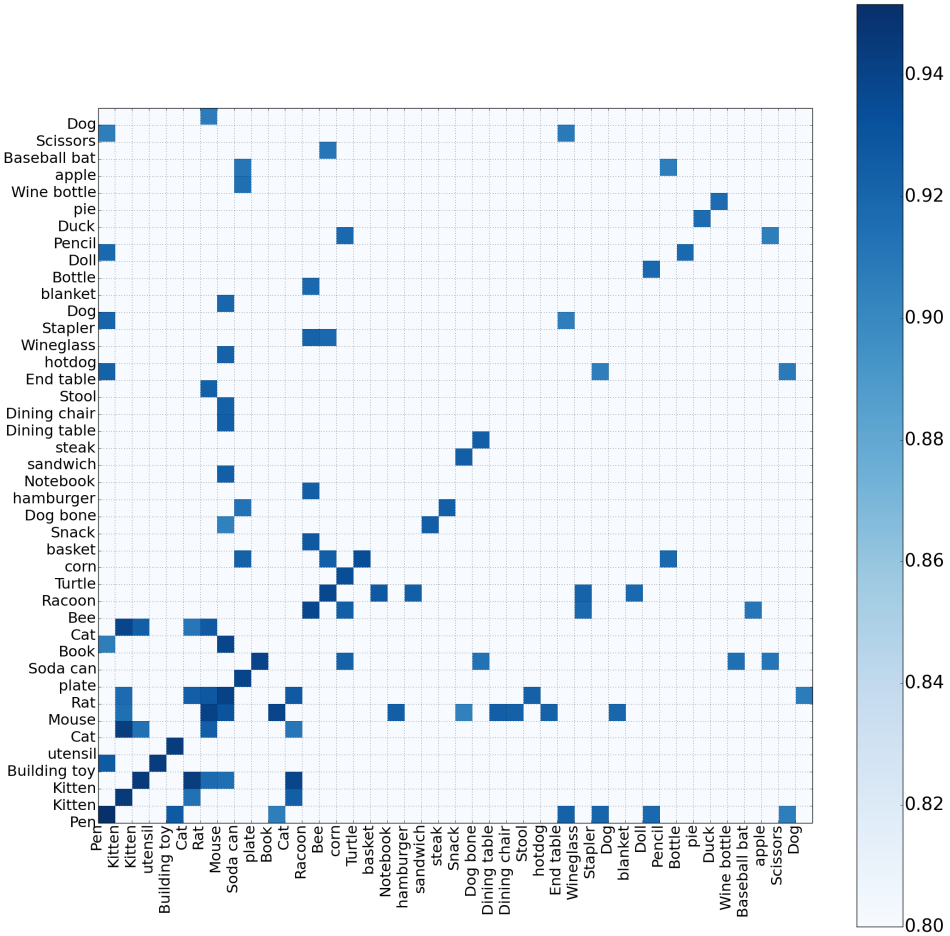



**Figure 41:** Top 100 object pairs that have the highest probabilities of cooccurring in a funny scene. Please note that repeated entries for an object type (*e.g.*, `dog`), correspond to slightly different versions (*e.g.*, breeds) of the same object type.

We observe that the humans that most appear in funny scenes are elderly people. This is probably because a number of scenes in our dataset depict old men behaving unexpectedly, *e.g.*, dancing or playing in the park as shown in Fig. 39(c), which is funny. Interestingly, we also observe that in general, animals appear more frequently in funny scenes. Animals like `'mouse`, `rat`, `raccoon` and `bee` appear in funny scenes significantly more than they do in unfunny scenes. Other objects having a strong bias towards appearing in funny `beehive`. Thus, we see that certain object types have a tendency to appear in funny scenes. A possible reason for this is that these objects are involved in funny interactions, or are intrinsically funny, and hence contribute to humor in these scenes.

**Funny Cooccurrence Matrix.** We populate two object cooccurrence matrices – **F** and **U**, corresponding to funny scenes and unfunny scenes, respectively. Each element in **F** and **U** corresponds to the count of the cooccurrence of a pair of objects across all funny and unfunny scenes, respectively. To enable the study of types of cooccurrences that contribute to humor, we compute the probability of a scene being funny, given that a pair of objects cooccur in the scene as $\frac{\mathbf{F}}{\mathbf{F}+\mathbf{U}}$, which is shown in Fig. 41 for the top 100 probable combinations that exist in a funny scene. Please note that repeated entries for an object type (*e.g.*, `dog`), correspond to slightly different versions (*e.g.*, breeds) of the same object type. An interesting set of object pairs that are present in funny scenes are `rat` appearing alongside `kitten`, `cat`, `stool`, and `dog`. Another interesting set of combinations is `raccoon` cooccurring with `bee`, `hamburger`, `basket`, and `wine glass`. We observe that this matrix captures interesting and unusual combinations of objects that appear together frequently in funny scenes.

## *A.6 User Interfaces*

In this section, we present the user interfaces that were used to collect data from AMT. Fig. 43 shows a screenshot of the user interface that we used to collect funny

scenes. Objects in the clipart library (on the *right* in the screenshot) can be dragged on to any part of the empty canvas shown in the figure. The pose, flip (*i.e.*, lateral orientation), and size of all objects can be changed once they are placed in the scene. In the case of humans, one of 8 expressions must be chosen (initially humans have blank faces) and fine-grained pose adjustments are required.

Fig. 44 shows the interface that we used to collect 'object-replaced' scenes for our FOR dataset. We showed workers an originally funny scene and asked them to replace objects in that scene so that the scene is not funny anymore. On clicking an object in the original scene, the object gets highlighted in green. A replacer object can then be chosen from the clipart library (displayed on the *right* in the screenshot). Objects that are replaced in the original scene show up in the empty canvas below. At any point, to undo a replacement, a user can click on the object in the below canvas and the corresponding object will be placed at its original position in the scene. The interface does not allow for the movement or the removal of objects.

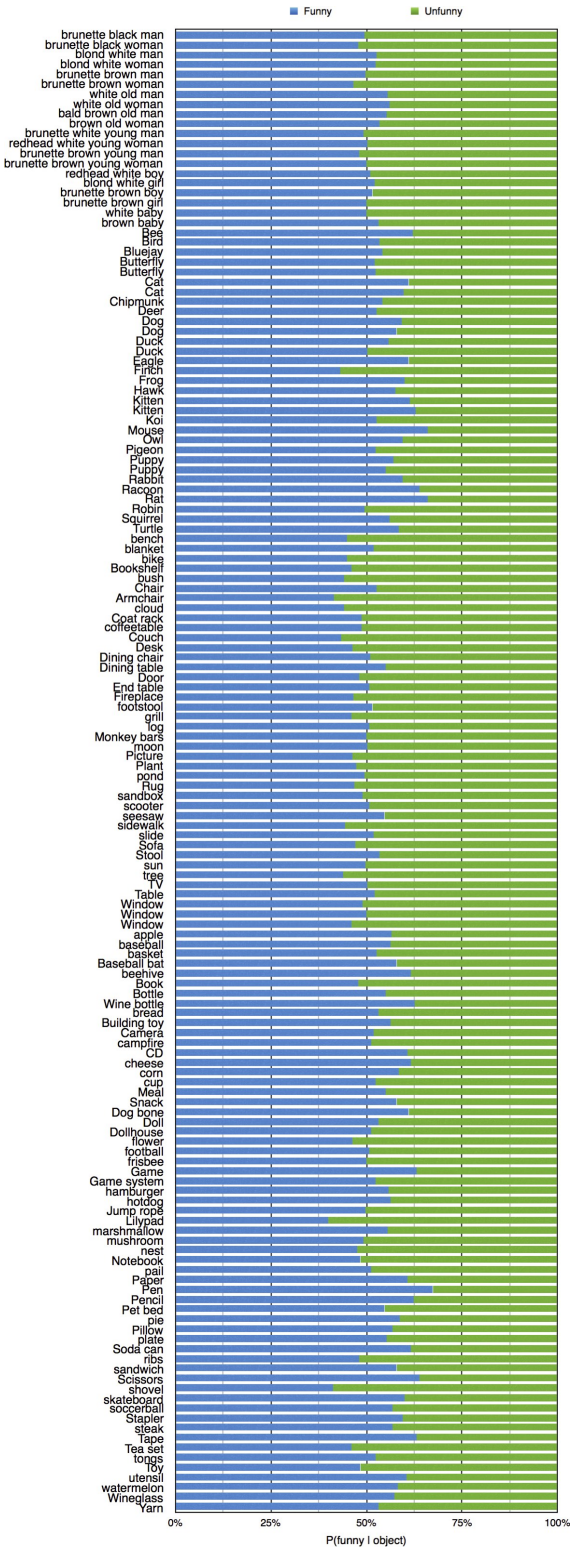



**Figure 42:** Probability of scene being funny, given object.

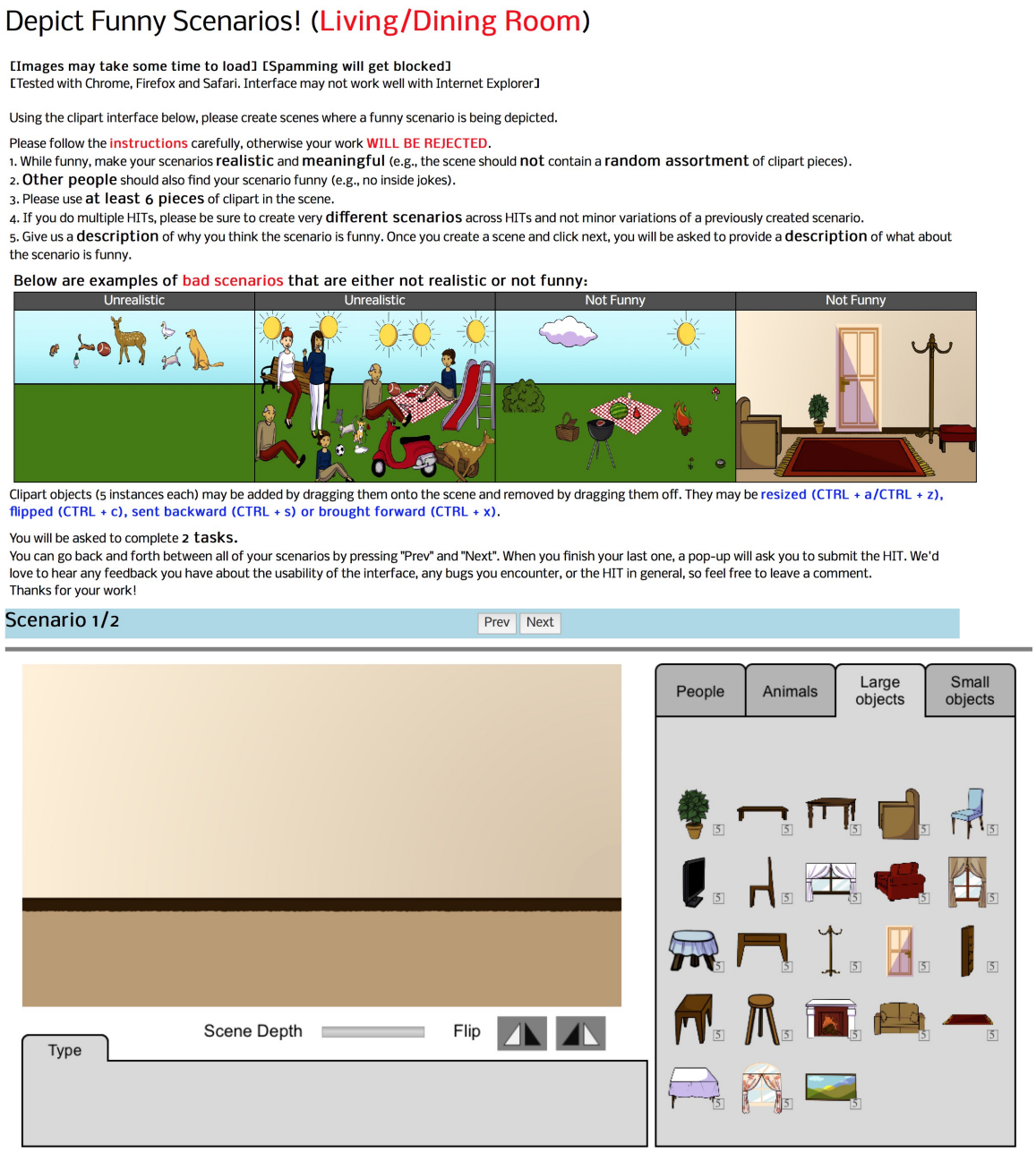



**Figure 43:** User interface used to create the funny scenes in the AVH dataset.

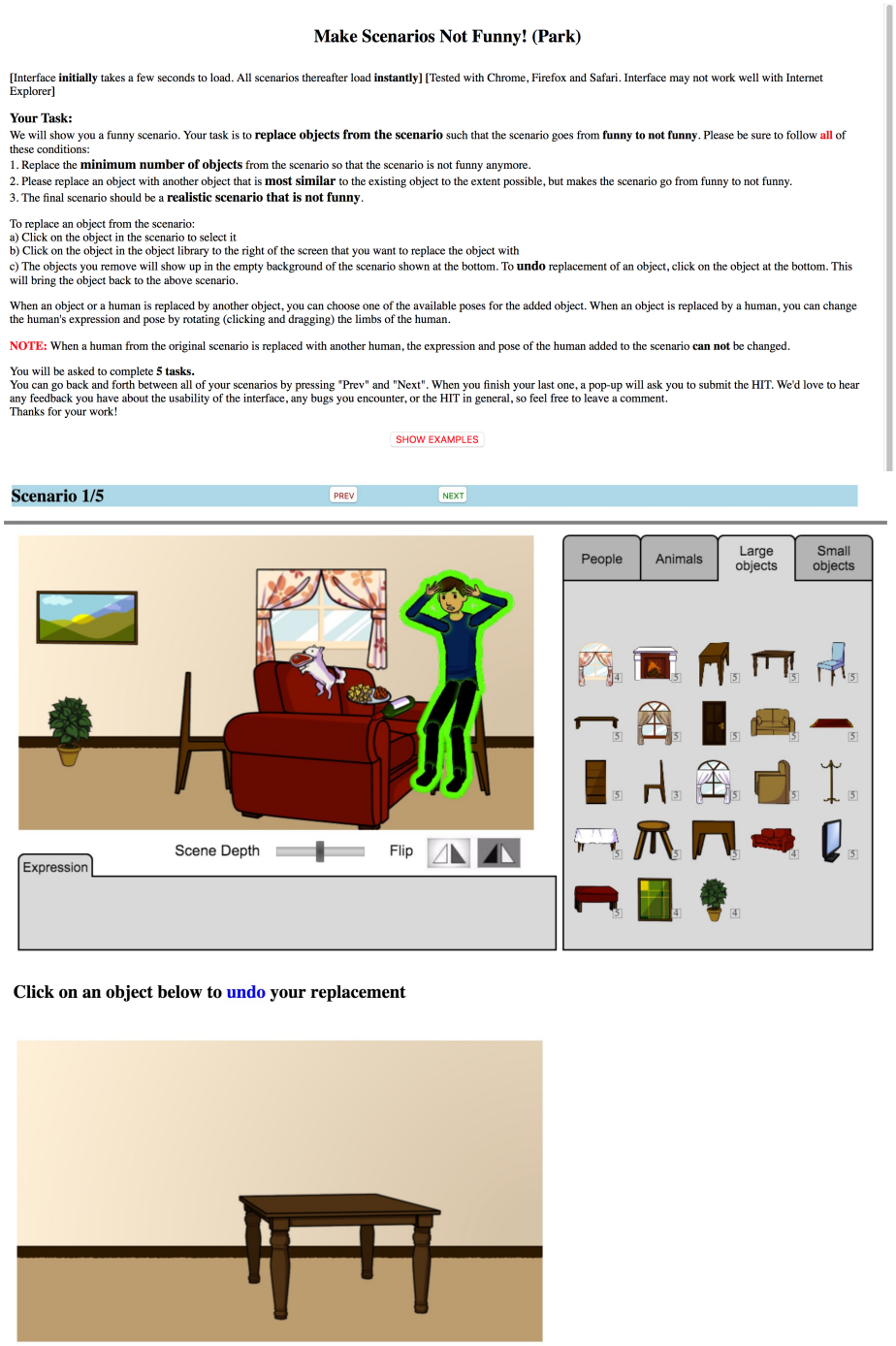



**Figure 44:** User interface to replace objects for the FOR dataset.

# APPENDIX B

# MULTI-MODAL HUMOR

## *B.1 Additional Experiments*

### B.1.1 Relevance of witty caption to image

We compared the relative relevance of the top witty caption from our generation approach against a machine generated boring caption (either for the same image or for a different, randomly chosen image) in a pairwise comparison. We showed Turkers an image and a pair of captions, and asked them to choose the more relevant caption for the image. We see that on average, the generated witty caption is considered more relevant than a machine generated boring caption for the same image 37.5% of the time. People found the generated witty caption to be more relevant than a random caption 97.2% of the time. This shows that in an effort to generate witty content, our approach produces descriptions that are a little less relevant compared to a boring description for the image. But our witty caption is clearly still relevant to the image (almost always more relevant than an unrelated caption).

### B.1.2 Retrieved captions vs. baselines

Humans evaluate the wittiness of each of the 3 top-ranked retrieved captions against baseline approaches and a human witty caption. As we see in Fig. 45, at $K = 1$, the top retrieved description is found to be wittier than only a human-written witty caption that is mismatched with the given image (witty mismatch) 83.8% of the time. The top retrieved caption is found *less* witty than even a typical caption (regular inference) about 63.4% of the time. Similarly, the retrieved caption is also found to be less witty than a naive method that produces punny captions (ambiguous) about 62% of the time. We observe the trend that as $K$ increases, recall also increases. On

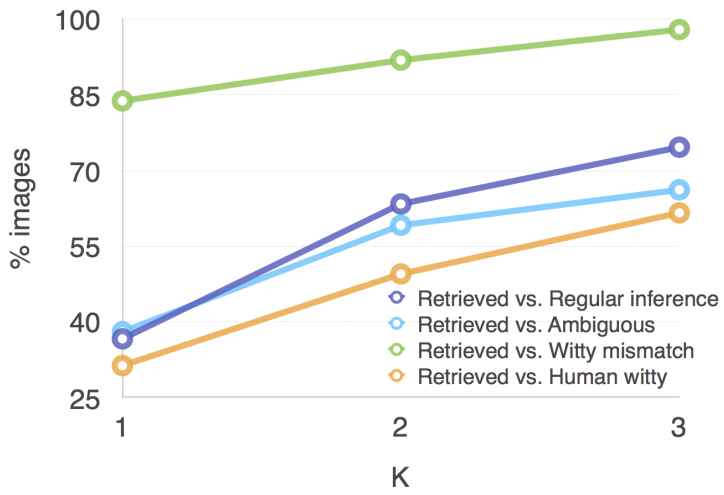



**Figure 45:** Comparison of wittiness of the top 3 captions from our retrieval approach vs. other approaches. The y-axis measures the % images for which at least one of $K$ captions from our approach is rated wittier than other approaches. As we increase the number of retrieved captions ($K$), recall steadily increases.

average, at least one of the top 3 retrieved captions is wittier than the (constrained) human witty caption about 61.6% of the time, compared to generated captions which are wittier 84.0% of the time.

Poor performance of retrieved captions could be due to the fact that they are often not perfectly apt for the given image since they are retrieved from story-based corpora. Please see Sec. 4.2 for examples, and a more detailed discussion. As we will see in the next section, these issues do not extend to the generation approach which exhibits strong performance against baseline approaches, human-written witty captions and the retrieval approach. While these captions might evoke a sense of incongruity, it is likely hard for the viewer to resolve the alternate interpretation of the retrieved caption as being applicable to the image.

## *B.2 Design choices*

In this section, we describe how our architecture design and parameter choices in the architectures influence witty descriptions. During the design of our model, we made choices of parameters based on observations from qualitative results. For instance,

we experimented with different beam sizes to generate a set of high precision captions with few false positives. We found that a beam size of 6 resulted in a sufficient number of candidate sentences which were reasonably accurate. We extract image tags from the top-K predictions of an image classifier. We experimented with different values of K, where $K \in \{1, 5, 10\}$. We also tried using a score threshold, where classes predicted with a score above the threshold were considered valid image tags. We found that $K = 5$ results in reasonable predictions. Determining a reasonable threshold on the other hand was difficult because for most images, class prediction scores are extremely peaky. We also experimented with the different positions that a pun counterpart can be forced to appear in. Based on qualitative examples, we found that the model generated witty descriptions that were somewhat sensible when a pun word appeared at any of the first or last 5 positions of a sentence. We also experimented with a number of different methods to re-rank the candidate of witty captions, e.g., language model score [104], image-sentence similarity score [117], semantic similarity (using Word2Vec [150]) of the pun counterpart to the sentence, a priori probability of the pun counterpart in a large corpus of English sentences to avoid rare / unfamiliar words, likelihood of the tag (under the image captioning model or the classifier as applicable). etc. We qualitatively found that re-ranking using log. prob. score of the image captioning model, while being the simplest, resulted in the best set of candidate witty captions.

## *B.3 Pun List*

Recall that we construct a list of puns by mining the web and based on automatic methods that measure the similarity of pronunciation of words. Upon inspecting our list of puns, we observe that it contains puns of many frequently used words and some pun words that are rarely used in everyday language, e.g., 'wight' (which is the counterpart of 'white'). Since a rare pun word can be distracting to a perceiver, the

corresponding caption might be harder to resolve, making it less likely to be perceived as witty. Thus, we see limited benefit in increasing the size of our pun list further to include words that are used even less frequently.

## *B.4 Interface for 'Be Witty!'*

We ask people on Amazon Mechanical Turk (AMT) to create witty descriptions for the given image. We also ask them to utilize one of the given pun words associated with the image. We show them a few good and bad examples to illustrate the task better. Fig. 46 shows the interface that we used to collect these human-written witty descriptions for an image.

## *B.5 Interface for 'Which is wittier?'*

We showed people on AMT two descriptions for a given image and asked them to click on the description that was wittier for the image. The web interface that we used to collect this data is shown in Fig. 47.

## *B.6 Sample images and witty descriptions*

The full set of 100 examples can be found on the author's webpage. Each image is accompanied by 4 witty descriptions from our generative and retrieval models – 3 top-ranked descriptions, and 1 low-ranked bad? description. We also provide the descriptions produced by the 3 baseline approaches – Regular inference, Witty mismatch and Ambiguous.

Hi, my name is **(F)Punky**. I am an **Artificial Intelligence (AI)**.
I am learning how to be witty. **Please write a caption about this image that contains a pun.**

**Keyboard shortcuts**

| | |
|---|---|
| **Previous** | Left arrow |
| **Next** | Right arrow |

**Task:** Write a witty sentence about the image containing one of the puns listed beside the image.

Please see a few good examples (green font) and bad examples (red font) below.

HIDE EXAMPLES

**Good examples**

**Witty sentence with a pun:**
Emotional wedding where the cake is in **tiers**.

**Witty sentence with a pun:**
A woman at a dine and **whine**.


**Witty sentence with a pun:**
A cat is pressing **pause** on the phone.

**Bad examples**

A bridesmaid is in **tiers** at a wedding.
**[Pun word should make sense! This caption makes sense for the *original* word but not for the pun.]**

She will always **whine** after wine.
**[No personal viewpoints. No first person accounts!]**


Sleepy cat said, "Dance to the music without pause".
**[Shouldn't be what a character in the picture might say!]**

**If you don't follow these instructions, your work will be rejected.**

PREVIOUS NEXT

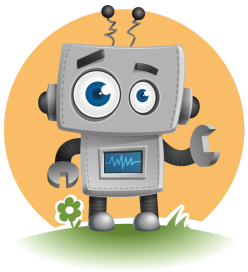
**Task 1/5**

**List of puns:** waul (wall), wight (white), stile (style), poll (pole), sine (sign)

Write a caption about this image using waul, wight, stile, poll, or sine.
Remember: The caption should be relevant to the image, and the sentence should make sense for: waul, wight, stile, poll, or sine.

Witty sentence here ...

**Figure 46:** AMT web interface for the 'Be Witty!' task.



Hi, my name is **(F)Punky**. I am an **Artificial Intelligence (AI)**.
I'm trying to learn to be **witty by using puns** while describing images. I'm not very good yet, and I'd like to learn so I can slowly get better.

Please tell me which of the following two captions are wittier for this image. To give you a sense for what pun I was going for -- I'll also show you in parenthesis what I saw in the image which I then made a pun around.

Even if both captions seem not all that witty, please indicate the one that seems (ever so slightly) wittier.

I will benefit from this positive feedback! Thanks :)

PREVIOUS **Task 5/15** NEXT

**Which of the two captions for the image is wittier?**

CAUGHT (COURT) A TENNIS PLAYER HITTING THE BALL .

THE TEDDY BEAR WERE BARE (BEAR)

**Keyboard shortcuts**

| | |
|---|---|
| **Top caption** | `Ctrl+j` |
| **Bottom caption** | `Ctrl+k` |
| **Previous** | `Ctrl+d` |
| **Next** | `Ctrl+h` |


**Figure 47:** AMT web interface for the 'Which is wittier?' task.

# APPENDIX C

# PREDICTABILITY

## *C.1 Introduction*

The appendix material is organized as follows. We first present further details regarding the role of familiarization which was presented in Sec. 6.2.1 of Chapter 6. Following this, in Sec. C.3, we discuss some of the challenges associated with data collection. In Sec. C.4, we provide more details about and discuss the automatic approaches (described in Sec. 6.2.2) used for Failure Prediction (FP) following ALERT [251]. Then, in Sec. C.5, we discuss various visual recognition scenarios in which a human might rely on an AI, and motivate the need for building a model of the AI in such scenarios. We then provide qualitative examples of montages that highlight the quirks which make VQA model predictable, and additionally share insights regarding the model from subjects who completed the tasks, in Sec. C.6. Finally, we describe an AMT survey we conducted to gauge public perception of AI, and provide a list of questions and qualitative analyses of results.

## *C.2 Analysis of the Role of Familiarization*

In Sec. 6.2.1, we presented human accuracies on the tasks of FP and KP. In this section, we draw the reader's attention towards more observations regarding familiarization of the subjects with the model. Following Sec. 6.2.1, the 29.84% improvement of IF over No IF for KP is significantly larger than that for FP (13.09%). This is understandable because a priori (No IF), KP is a much harder task as compared to FP due to the increased space of possible subject responses given a QI pair, and the combination of relevant and irrelevant QI-pairs in the test phase.

**VQA Researchers.** Just as an anecdotal point of reference, we also conducted experiments across experts with varying degrees of familiarity with agents like Vicki. We observed that a VQA researcher had an accuracy of 80% versus a computer vision (but not VQA) researcher who had 60% in a shorter version of the FP task without instant feedback. Clearly, familiarity with Vicki appears to play a critical role in how well a human can predict its oncoming failures or successes.

## *C.3 Data Collection*

We describe the setup of the experiments we performed in Sec. 6.2.1. Recall that we conduct our studies on Amazon Mechanical Turk. In this section, we describe the challenges involved in collecting data for the set of experiments we described. Collecting data for our setup is challenging because: (1) Each subject undergoes 'training' to become familiar with Vicki before they are tested. This results in the AMT tasks being unusually long (mean HIT durations across the tasks of FP and KP = $10.11 \pm 1.09$ and $24.49 \pm 1.85$ min. respectively). Crucially, this also reduces the subject pool to only those willing to participate in long tasks. (2) Once a subject does one collaborative task with Vicki, they cannot do another because the training / familiarity with Vicki would leak over. This constraint causes our analyses to require as many subjects as tasks. Since work division in crowdsourcing tasks follows a Pareto principle [127], this makes data collection very slow. In light of these challenges, we focus on a small set of questions to systematically evaluate the role of training and exposure to Vicki's internal states.

## *C.4 Automatic Approaches for FP*

In order to put human accuracies on FP in perspective, we also evaluate automatic approaches (described in Sec. 6.2.2) that determine Vicki's failure or success from its internal states. We train a Multilayer Perceptron (MLP) on Vicki's output scores (post-softmax) to predict success vs failure. This achieves an FP accuracy of 81%.

Training an MLP which takes as input question features (average word2vec embeddings [150] of words in the question) concatenated with image features (fc7 from VGG-19) to predict success vs failure (which we call ALERT following [251]) achieves an FP accuracy of 65%. These methods are trained on about 66% of the VQA 1.0-val set (∼81k examples, rest used for validation), while human subjects are trained on only 50 examples. Note that we only report machine results to put human accuracies in perspective. We do not draw any inferences about the relative capabilities of both.

## *C.5 Visual Recognition Scenarios*

In general, one might wonder why a human would need Vicki to answer questions if they are already looking at the image. This may be true for the VQA dataset, but outside of that there are scenarios where the human either does not know the answer to a question of interest (e.g., the species of a bird), or the amount of visual data is so large (e.g., long surveillance videos) that it would be prohibitively cumbersome for them to sift through it. Note that even in this scenario where the human does not know the answer to the question, a human who understands Vicki's failure modes from past experience would know when to trust its decision. For instance, if the bird is occluded, or the scene is cluttered, or the lighting is bad, or the bird pose is odd, Vicki will likely fail. Moreover, the idea of humans predicting the AI's failure also applies to other scenarios where the human may not be looking at the image, and hence needs to work with Vicki (e.g., blind user, or a human working with a teleoperated robot). In these cases too, it would be useful for the human to have a sense for the contexts and environments and/or kinds of questions for which Vicki can be trusted. In this work, as a first step, we focus on the first scenario where the human is looking at the image and a question while predicting Vicki's failures and responses.

## *C.6 Vicki's Quirks*

We present some examples in Fig. 48 and Fig. 49 that highlight Vicki's quirks. Recall that there are several factors which lead to Vicki being quirky, many of which are well known in VQA literature [1]. As we can see across both examples, Vicki exhibits these quirks in a somewhat predictable fashion. At first glance, the primary factors that seem to decide Vicki's response to a question given an image are the properties and activities associated with the salient objects in the image, in combination with the language and the phrasing of the question being asked. This is evident when we look across the images (see Fig. 48 and 49) for question-answer (QA) pairs such as – *What are the people doing? Grazing*, *What is the man holding? Cow* and *Is it raining? No.* As a specific example, notice the images for the QA pair *What color is the grass? Blue* (see Fig. 48) – Vicki's response to this question is the most dominant color in the scene across all images even though there is no grass present in any of them. Similarly, for the QA pair *What does the sign say? Banana* (see Fig. 49) – Vicki's answer is the salient object across all the scenes.

Interestingly, some subjects did try and pick up on some of the quirks and beliefs described previously, and formed a mental model of Vicki while completing the Failure Prediction or Knowledge Prediction tasks. We asked subjects to leave comments after completing a task and some of them shared their views on Vicki's behavior. We share some of those comments below. The abbreviations used are Failure Prediction (FP), Knowledge Prediction (KP) and Instant Feedback (IF).

1. **FP**
    - *These images were all pretty easy to see what animal it was. I would imagine the robot would be able to get 90% of the animals correct, unless there were multiple animals in the same photo.*

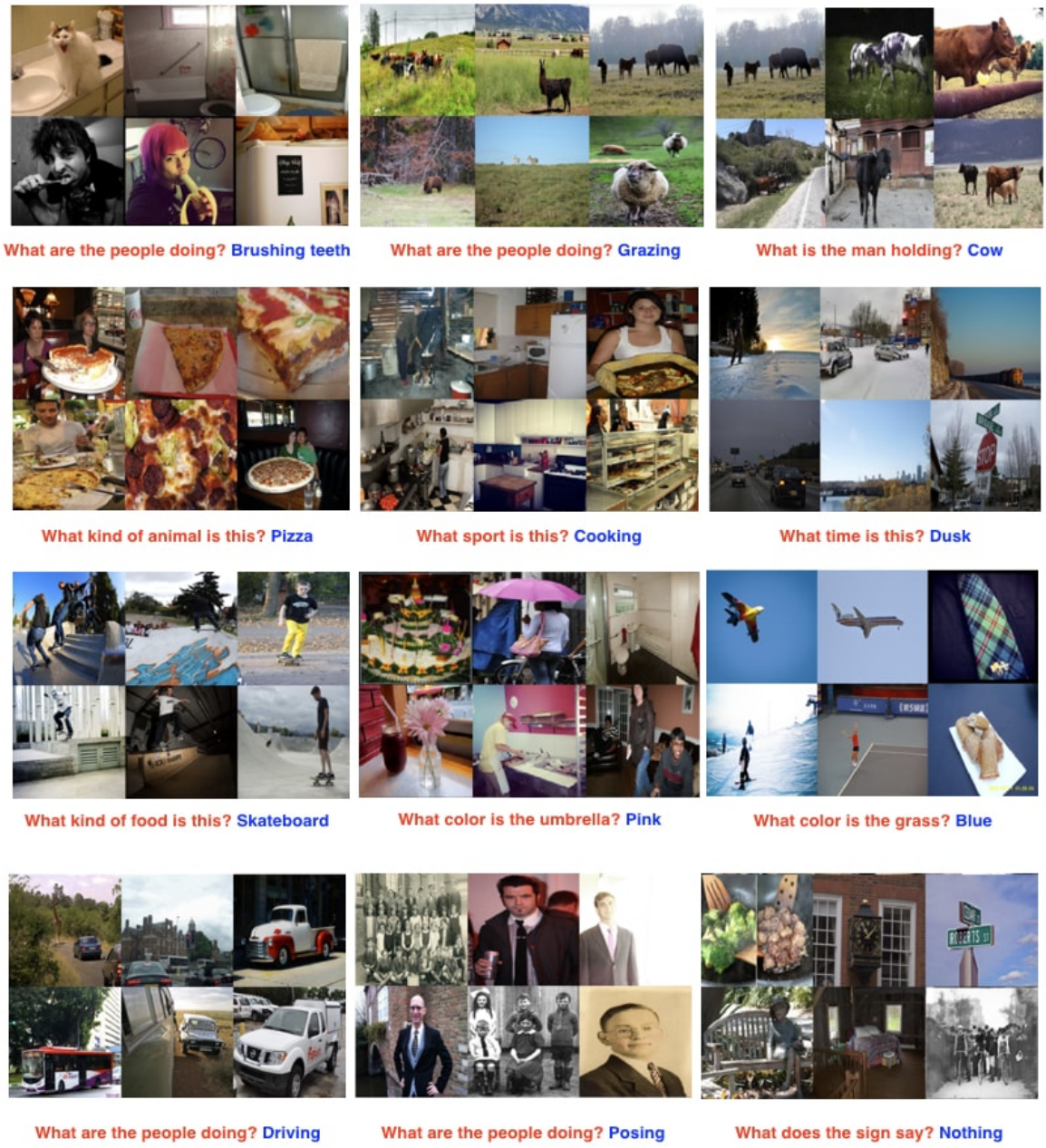



**Figure 48:** Given a question (red) we show images for which Vicki gave the same answer (blue) to the question to observe Vicki's quirks.


**Figure 49:** Given a question (red) we show images for which Vicki gave the same answer (blue) to the question to observe Vicki's quirks.


**Figure 50:** We show a word cloud of all the comments left by subjects after completing the tasks across all settings. From the frequency of positive comments about the tasks, it appears that subjects were enthusiastic to familiarize themselves with Vicki.

- *I think the brighter the color the more likely they are to get it right. Multi-colored, not so sure.*
- *I'd love to know the answers to these myself.*

2. **FP + IF**

- *This is fun, but kind of hard to tell what the hints mean. Can she determine the color differences in multi-colored umbrellas or are they automatically marked wrong because she only chooses one color instead of all of the colors? It seems to me that she just goes for the brightest color in the pic. This is very interesting. Thank you! :)*
- *I didn't quite grasp what the AI's algorithm was for determining right or wrong. I want to say that it was if the AI could see the face of the animal then it guessed correctly, but I'm really not sure.*

3. **FP + IF + Explanation Modalities**

    - *Even though Vicki is looking at the right spot doesn't always mean she will guess correctly. To me there was no rhyme or reason to guessing correctly. Thank you.*
    - *I think she can accurately know a small number of people but cannot know a huge grouping yet.*
    - *I would be more interested to find out how Vickis metrics work. What I was assuming is just color phase and distance might not be accurate.*

4. **KP**

    - *Time questions are tricky because all Vicki can do is round to the nearest number.*
    - *there were a few that seemed like it was missing obvious answers - like bus and bus stop but not bus station. Also words like lobby seemed to be missing.*

5. **KP + IF**

    - *Interesting, though it seems Vicki has a lot more learning to do. Thank you!*
    - *This HIT was interesting, but a bit hard. Thank you for the opportunity to work this.*

6. **KP + IF + Explanation Modalities**

    - *You need to eliminate the nuances of night time and daytime from the computer and choose one phrasing "night" or "day" Vicki understands. The nuance keeps me and I'm sure others obtaining a higher score here on this task.*

- *I felt that Vickie was mistaken as to what some colors were for the first test which probably carried over and I tried my best to recreate her responses.*

7. **KP + IF + Montages**

   - *I am not sure that I ever completely understood how Vicki thought. It seemed it had more to do with what was in the pictures instead of the time of day it looked in the pictures. If there was food, she chose noon or morning, even though at times it was clearly breakfast food and she labeled it noon.*
   - *It doesn't seem very accurate as I made sure to count and took my time assessing the pictures.*
   - *it is hard to figure out what they are looking for since there isn't many umbrellas in the pictures*

On a high-level reading through all comments, we found that subjects felt that Vicki's response often revolves around the most salient object in the image, that Vicki is bad at counting, and that Vicki often responds with the most dominant color in the image when asked a color question. In Fig. 51(a), we show a word cloud of all the comments left by the subjects after completing the tasks. From the comments, we observed that subjects were very enthusiastic to familiarize themselves with Vicki, and found the process engaging. Many thought that the scenarios presented to them were interesting and fun, despite being hard. We used some basic elements of gamification, such as performance-based reward and narrative, to make our tasks more engaging; we think the positive response indicates the possibility of making such human-familiarization with AI engaging even in real-world settings.

## *C.7 Perception of AI*

In addition to measuring the subjects' capabilities to predict Vicki's behavior, we also conducted a survey to assess their general impressions of present-day AI. Specifically,

we asked them to fill out a survey with questions focusing around three types of information - "Background Information", "Familiarity with Computers and AI" and "Estimates of AI's capabilities".

In Fig. 52, 53 and 54, we break down the 321 subjects that completed the survey by their response to each question.

As part of the survey, subjects were also asked a few subjective questions about their opinions on present–day AI's capabilities. These include multiple-choice questions focusing on some specific capabilities of AI ("Can AI recognize faces?", "Can AI drive cars?", etc.) – responses to which are summarized in Fig. 54. The subjects were also asked to specifically list tasks that they thought AI is capable of performing *today* (see Fig. 51(b)), will be capable of *in the next 3 years* (see Fig. 51(c)), and will be capable of *in the next 10 years* (see Fig. 51(d)). We also asked how *they* think AI works (see Fig. 51(e)). In Fig. 51(b), 51(c) and 51(d), we show word clouds corresponding to what subjects thought about the capabilities of AI. We also share some of those responses below.

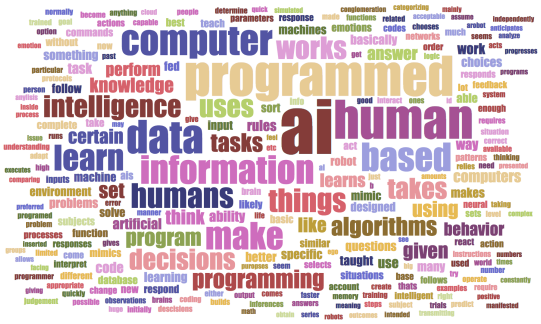
1. Name three things that you think AI today can do. *Predict sports games; Detect specific types of cancer in images; Control house temp based on outside weather; translate; calculate probabilities; Predictive Analysis; AI can predict future events that happen like potential car accidents; lip reading; code; Facial recognition; Drive cars; Play Go; predict the weather; Hold a conversation; Be a personal assistant; Speech recognition; search the web quicker.*
2. Name three things that you think AI today can't yet do but will be able to do in 3 years. *Fly planes; Judge emotion in voices; Predict what I want for dinner; perform surgery; drive cars; manage larger amounts of information at a faster rate; think independently totally; play baseball; drive semi trucks; Be a caregiver; anticipate a person's lying ability; read minds; Diagnose patients;*


(a) We show a word cloud of all the comments left by subjects after completing the tasks across all settings. From the frequency of positive comments about the tasks, it appears that subjects were enthusiastic to familiarize themselves with Vicki.


(b) A word cloud of subject responses to "Name three things that you think AI today can do."


(c) A word cloud of subject responses to "Name three things that you think AI today can't yet do but will be able to do in 3 years"


(d) A word cloud of subject responses to "Name three things that you think AI today can't yet do and will take a while ($>$ 10 years) before it can do it."




(e) A word cloud of subject responses when asked to describe how *they* think AI today works.

**Figure 51:** Word clouds corresponding to responses from humans for different questions.

*improve robots to walk straight; Run websites; solve complex problems like climate change issues; program other ai; guess ages; form conclusions based on evidence; act on more complex commands; create art.*

3. Name three things that you think AI today can't yet do and will take a while (> 10 years) before it can do it. *Imitate humans; be indistinguishable from humans; read minds; Have emotions; Develop feelings; make robots act like humans; truly learn and think; Replace humans; impersonate people; teach; be a human; full AI with personalities; Run governments; be able to match a human entirely; take over the world; Pass a Turing test; be a human like friend; intimacy; Recognize things like sarcasm and humor.*

Interestingly, we observe a steady progression in subjects' expectations of AI's capabilities, as the time span increases. Reading through the responses, we notice that subjects believe that AI today can successfully perform tasks such as *machine translation*, *driving vehicles*, *speech recognition*, *analyzing information and drawing conclusions*, etc. (see Fig. 51(b)). It is likely that this is influenced by the subjects' exposure to or interaction with some form of AI in their day-to-day lives. When asked about what AI can do three years from now, most subjects suggested more sophisticated tasks such as *inferring emotions from voice tone*, *performing surgery*, and even *dealing with climate change issues* (see Fig. 51(c)). However, the most interesting trends emerge while observing subjects' expectation of what AI can achieve in the next 10 years (see Fig. 51(d)). A major proportion of subjects believe that AI will gain the ability to *understand and emulate human beings*, *teach human beings*, *develop feelings and emotions* and *pass the Turing test*.

We also observe how subjects think AI works (see Fig. 51(e)). Mostly, subjects believe that an AI agent today is a system with high computational capabilities that has been programmed to simulate intelligence and perform certain tasks by exposing it to huge amounts of information, or, as one of subjects phrased it – *broadly AI*

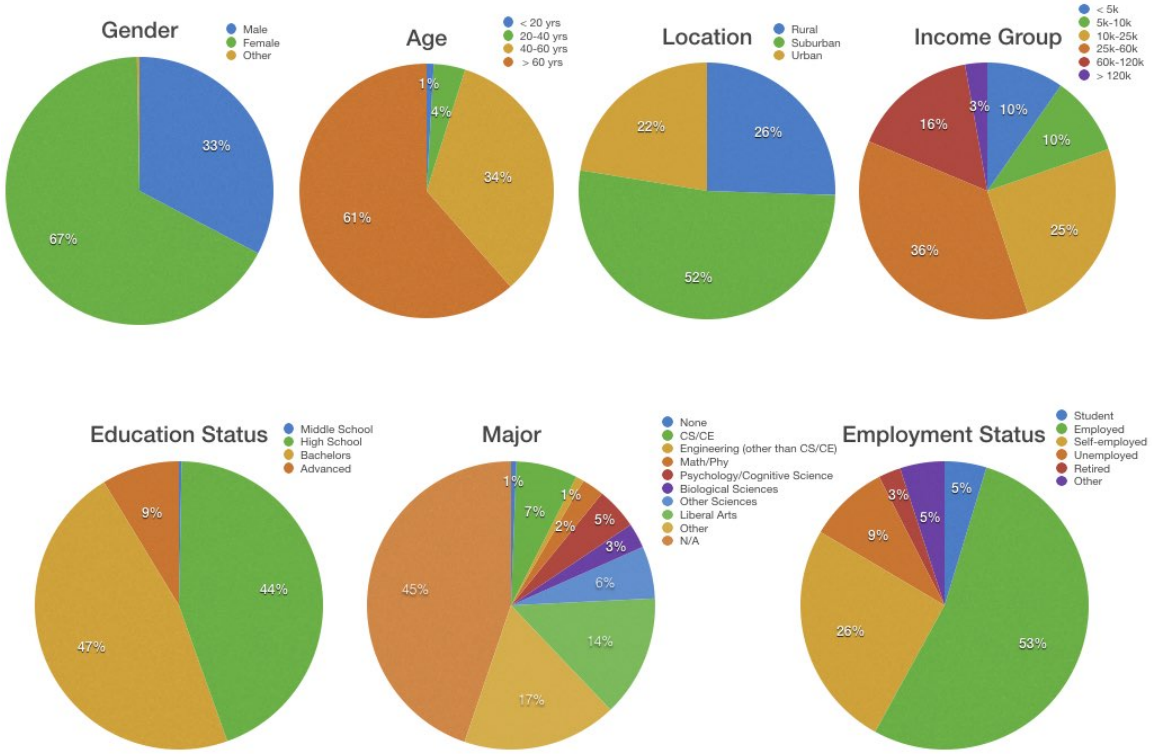



**Figure 52:** Population Demographics (across 321 subjects)

*recognizes patterns and creates optimal actions based on those patterns towards some predefined goals*. In summary, it appears that subjects have high expectations from AI, given enough time. While it is uncertain at this stage how many, or how soon, these feats will actually be achieved, we believe that building a model of the AI's skillset will help humans generally become more active and effective collaborators in human–AI teams.

We now provide a full list of questions the subjects were asked in the survey.

1. How old are you?
   (a) Less than 20 years
   (b) Between 20 and 40 years
   (c) Between 40 and 60 years
   (d) Greater than 60 years

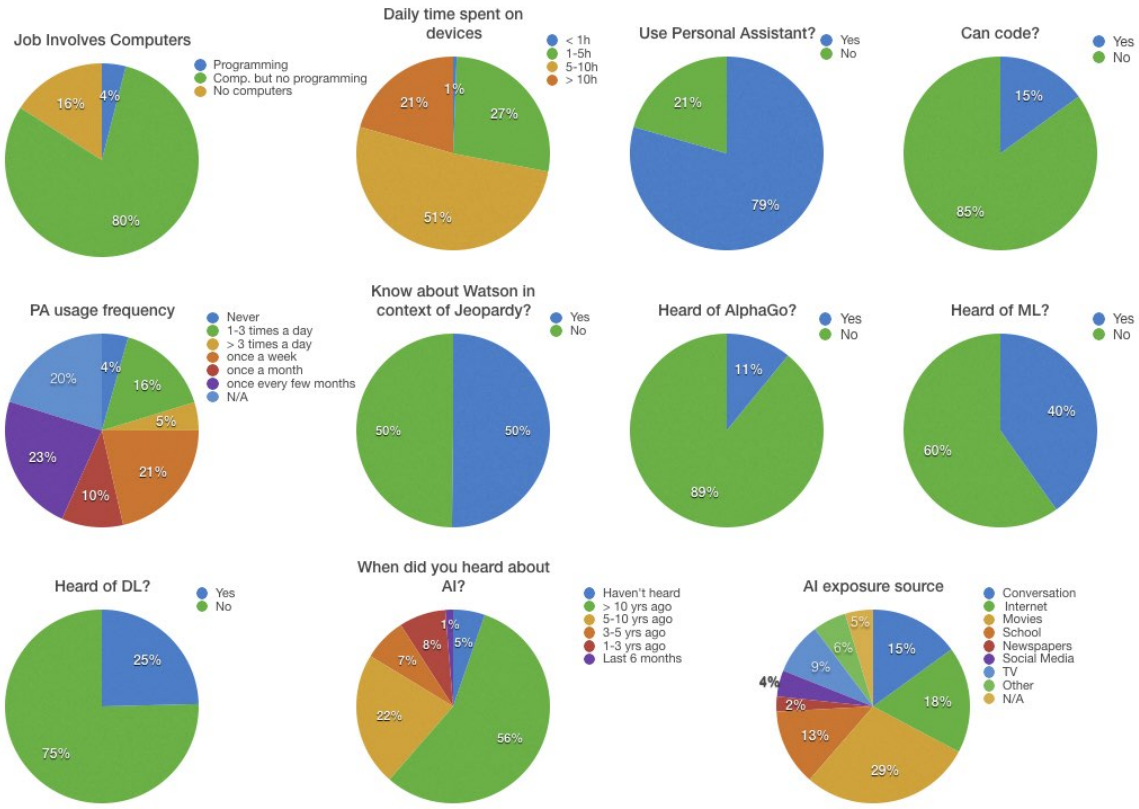



**Figure 53:** Technology and AI exposure (across 321 subjects)

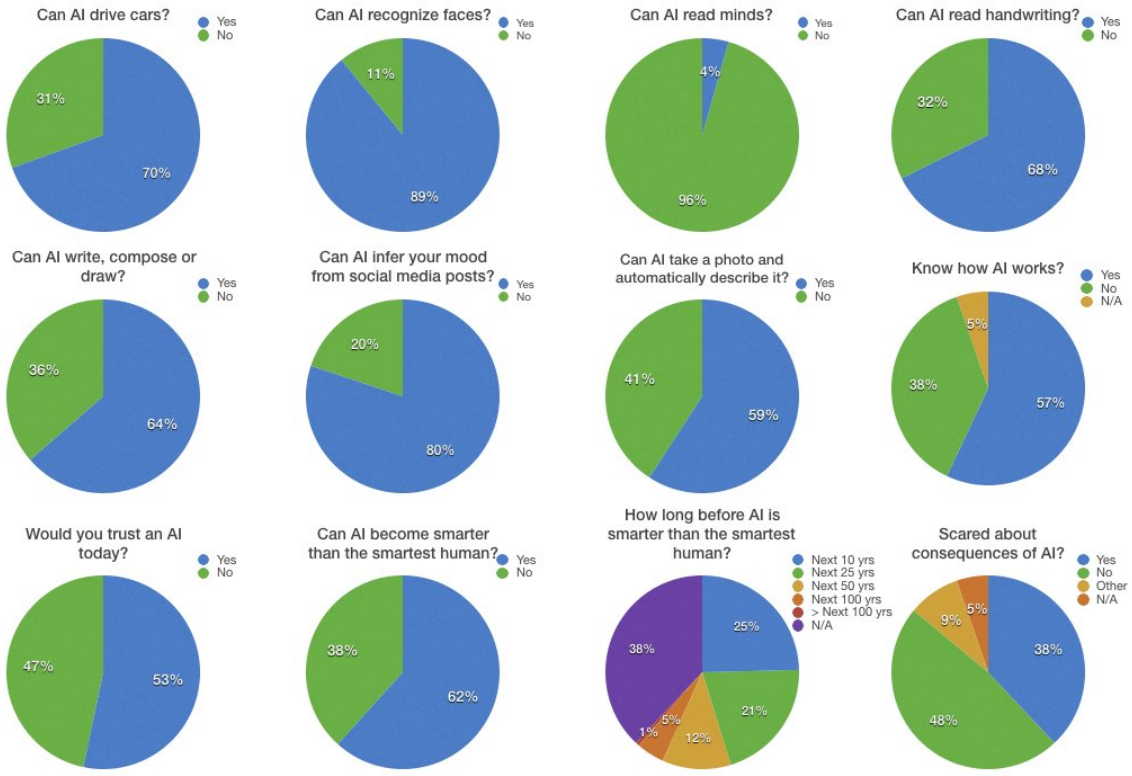



**Figure 54:** Perception of AI (across 321 subjects)

2. What is your gender?
    (a) Male
    (b) Female
    (c) Other
3. Where do you live?
    (a) Rural
    (b) Suburban
    (c) Urban
4. Are you?
    (a) A student
    (b) Employed
    (c) Self-employed
    (d) Unemployed
    (e) Retired
    (f) Other
5. To which income group do you belong?
    (a) Less than 5000$ per year
    (b) 5,000-10,000$ per year
    (c) 10,000-25,000$ per year
    (d) 25,000-60,000$ per year
    (e) 60,000-120,000$ per year
    (f) More than 120,000$ per year
6. What is your highest level of education?
    (a) No formal education
    (b) Middle School
    (c) High School
    (d) College (Bachelors)

(e) Advanced Degree

7. What was your major?
   (a) Computer Science / Computer Engineering
   (b) Engineering but not Computer Science
   (c) Mathematics / Physics
   (d) Philosophy
   (e) Biology / Physiology / Neurosciences
   (f) Psychology / Cognitive Sciences
   (g) Other Sciences
   (h) Liberal Arts
   (i) Other
   (j) None
8. Do you know how to program / code?
   (a) Yes
   (b) No
9. Does your full-time job involve:
   (a) No computers
   (b) Working with computers but no programming / coding?
   (c) Programming / Coding
10. How many hours a day do you spend on your computer / laptop / smartphone?
   (a) Less than 1 hour
   (b) 1-5 hours
   (c) 5-10 hours
   (d) Above 10 hours
11. Do you know what Watson is in the context of Jeopardy?
   (a) Yes
   (b) No

12. Have you ever used Siri, Alexa, or Google Now/Google Assistant?
    (a) Yes
    (b) No
13. How often do you use Siri, Alexa, Google Now, Google Assistant, or something equivalent?
    (a) About once every few months
    (b) About once a month
    (c) About once a week
    (d) About 1-3 times a day
    (e) More than 3 times a day
14. Have you heard of AlphaGo?
    (a) Yes
    (b) No
15. Have you heard of Machine Learning?
    (a) Yes
    (b) No
16. Have you heard of Deep Learning?
    (a) Yes
    (b) No
17. When did you first hear of Artificial Intelligence (AI)?
    (a) I have not heard of AI
    (b) More than 10 years ago
    (c) 5-10 years ago
    (d) 3-5 years ago
    (e) 1-3 years ago
    (f) In the last six months
    (g) Last month

18. How did you learn about AI?
    (a) School / College
    (b) Conversation with people
    (c) Movies
    (d) Newspapers
    (e) Social media
    (f) Internet
    (g) TV
    (h) Other
19. Do you think AI today can drive cars fully autonomously?
    (a) Yes
    (b) No
20. Do you think AI today can automatically recognize faces in a photo?
    (a) Yes
    (b) No
21. Do you think AI today can read your mind?
    (a) Yes
    (b) No
22. Do you think AI today can automatically read your handwriting?
    (a) Yes
    (b) No
23. Do you think AI today can write poems, compose music, make paintings?
    (a) Yes
    (b) No
24. Do you think AI today can read your Tweets, Facebook posts, etc. and figure out if you are having a good day or not?
    (a) Yes

(b) No

25. Do you think AI today can take a photo and automatically describe it in a sentence?
    (a) Yes
    (b) No
26. Other than those mentioned above, name three things that you think AI today can do.
27. Other than those mentioned above, name three things that you think AI today can't yet do but will be able to do in 3 years.
28. Other than those mentioned above, name three things that you think AI today can't yet do and will take a while ($>$ 10 years) before it can do it.
29. Do you have a sense of how AI works?
    (a) Yes
    (b) No
    (c) If yes, describe in a sentence or two how AI works.
30. Would you trust an AI's decisions today?
    (a) Yes
    (b) No
31. Do you think AI can ever become smarter than the smartest human?
    (a) Yes
    (b) No
32. If yes, in how many years?
    (a) Within the next 10 years
    (b) Within the next 25 years
    (c) Within the next 50 years
    (d) Within the next 100 years
    (e) In more than 100 years

33. Are you scared about the consequences of AI?

    (a) Yes

    (b) No

    (c) Other

    (d) If other, explain.

# APPENDIX D

# LIST OF PUBLICATIONS

1. We Are Humor Beings: Understanding and Predicting Visual Humor.

   **A. Chandrasekaran**, A. Kalyan, S. Antol, M. Bansal, D. Batra, C. L. Zitnick, D. Parikh.

   Spotlight at CVPR (Conference in Computer Vision and Pattern Recognition), 2016.

2. Sort Story: Sorting Jumbled Images and Captions into Stories.

   H. Agrawal*, **A. Chandrasekaran***, D. Batra, D. Parikh, M. Bansal.

   Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016.

3. Evaluating Visual Conversational Agents via Cooperative Human-AI Games.

   P. Chattopadhyay*, D. Yadav*, V. Prabhu, **A. Chandrasekaran**, A. Das, S. Lee, D. Batra, D. Parikh.

   AAAI Conference on Human Computation and Crowdsourcing (HCOMP), 2017.

4. It Takes Two to Tango: Towards Theory of AI's Mind.

   **A. Chandrasekaran***, D. Yadav*, P. Chattopadhyay*, V. Prabhu*, D. Parikh.

   Chalearn Looking at People Workshop at CVPR, 2017.

5. Punny Captions: Witty Wordplay in Image Descriptions.

   **A. Chandrasekaran**, D. Parikh, M. Bansal.

   North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), 2018.

6. Do explanation modalities make VQA models more predictable to a human?
   **A. Chandrasekaran***, V. Prabhu*, D. Yadav*, P. Chattopadhyay*, D. Parikh. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.

* Denotes equal contribution.

# REFERENCES


[1] Agrawal, A., Batra, D., and Parikh, D., "Analyzing the behavior of visual question answering models," *arXiv preprint arXiv:1606.07356*, 2016.

[2] Agrawal, H., Chandrasekaran, A., Batra, D., Parikh, D., and Bansal, M., "Sort story: Sorting jumbled images and captions into stories," *arXiv preprint arXiv:1606.07493*, 2016.

[3] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.

[4] Andreas, J. and Klein, D., "Reasoning about pragmatics with neural listeners and speakers," *arXiv preprint arXiv:1604.00562*, 2016.

[5] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D., "VQA: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

[6] Antol, S., Zitnick, C. L., and Parikh, D., "Zero-Shot Learning via Visual Abstraction," in *ECCV*, 2014.

[7] Aristotle and McKeon, R., *The Basic Works of Aristotle*. Modern Library, 2001.

[8] Attardo, S., *Linguistic theories of humor*. Walter de Gruyter, 1994.

[9] Attardo, S. and Raskin, V., "Script theory revis (it) ed: Joke similarity and joke representation model," *Humor-International Journal of Humor Research*, vol. 4, no. 3-4, pp. 293–348, 1991.

[10] Ba, J., Mnih, V., and Kavukcuoglu, K., "Multiple object recognition with visual attention," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[11] Bahdanau, D., Cho, K., and Bengio, Y., "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[12] Bansal, A., Farhadi, A., and Parikh, D., "Towards transparent systems: Semantic characterization of failure modes," in *European Conference on Computer Vision*, pp. 366–381, Springer, 2014.

[13] Baron-Cohen, S., *The evolution of a theory of mind*. na, 1999.

[14] Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I., "The reading the mind in the eyes test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism," *Journal of child psychology and psychiatry*, vol. 42, no. 2, pp. 241–251, 2001.

[15] Basha, T., Moses, Y., and Avidan, S., "Photo sequencing," *International Journal of Computer Vision*, 2014.

[16] Berg, A. C., Berg, T. L., Daume, H., Dodge, J., Goyal, A., Han, X., Mensch, A., Mitchell, M., Sood, A., Stratos, K., and others, "Understanding and predicting importance in images," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3562–3569, IEEE, 2012.

[17] Berger, A. A., *An Anatomy of Humor*. New Brunswick, N.J., U.S.A.: Transaction Publishers, 1993.

[18] Bharata-Muni and Ghosh, M., "Natya shastra (with english translations)," 1951.

[19] Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., and others, "Vizwiz: nearly real-time answers to visual questions," in *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp. 333–342, ACM, 2010.

[20] Bilgic, M. and Mooney, R. J., "Explaining recommendations: Satisfaction vs. promotion," in *Beyond Personalization Workshop, IUI*, vol. 5, p. 153, 2005.

[21] Binsted, K., "Machine humour: An implemented model of puns," *PhD Thesis, University of Edinburgh*, 1996.

[22] Binsted, K., Bergen, B., O'Mara, D., Coulson, S., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., and Pain, H., "Computational humor," *IEEE Intelligent Systems*, 2006.

[23] Binsted, K. and Ritchie, G., "Computational rules for generating punning riddles," *Humor: International Journal of Humor Research*, 1997.

[24] Bippus, A. M., "Making sense of humor in young romantic relationships: Understanding partners' perceptions," *Humor: International Journal of Humor Research*, 2000.

[25] Bird, S., Klein, E., and Loper, E., *Natural Language Processing with Python*. O'Reilly Media, 2009.

[26] Boguraev, B. and Ando, R. K., "Timeml-compliant text analysis for temporal reasoning.," in *IJCAI*, 2005.

[27] Bornstein, A. M., "Is artificial intelligence permanently inscrutable?," 2016.

[28] Bosselut, A., Chen, J., Warren, D., Hajishirzi, H., and Choi, Y., "Learning prototypical event structure from photo albums," in *ACL*, 2016.

[29] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D., "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

[30] Bressler, E. R. and Balshine, S., "The influence of humor on desirability," *Evolution and Human Behavior*, 2006.

[31] Bressler, E. R., Martin, R. A., and Balshine, S., "Production and appreciation of humor as sexually selected traits," *Evolution and Human Behavior*, 2006.

[32] Broadbent, E., Kumar, V., Li, X., Sollers 3rd, J., Stafford, R. Q., MacDonald, B. A., and Wegner, D. M., "Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality," *PloS one*, vol. 8, no. 8, p. e72589, 2013.

[33] Buijzen, M. and Valkenburg, P. M., "Developing a typology of humor in audiovisual media," *Media Psychology*, 2004.

[34] Buss, D. M., "The evolution of human intrasexual competition: Tactics of mate attraction.," *Journal of Personality and Social Psychology*, 1988.

[35] Chambers, N. and Jurafsky, D., "Unsupervised learning of narrative event chains.," in *ACL*, Citeseer, 2008.

[36] Chambers, N., Wang, S., and Jurafsky, D., "Classifying temporal relations between events," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Association for Computational Linguistics, 2007.

[37] Chandra, A. K., Kozen, D. C., and Stockmeyer, L. J., "Alternation," *Journal of the Association for Computing Machinery*, vol. 28, no. 1, pp. 114–133, 1981.

[38] Chandrasekaran, A., Kalyan, A., Antol, S., Bansal, M., Batra, D., Zitnick, C. L., and Parikh, D., "We are humor beings: Understanding and predicting visual humor," in *CVPR*, 2016.

[39] Chandrasekaran, A., Vijayakumar, A. K., Antol, S., Bansal, M., Batra, D., Lawrence Zitnick, C., and Parikh, D., "We are humor beings: Understanding and predicting visual humor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4603–4612, 2016.

[40] Chattopadhyay, P., Vedantam, R., Ramprasaath, R. S., Batra, D., and Parikh, D., "Counting everyday objects in everyday scenes," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[41] Chattopadhyay, P., Yadav, D., Prabhu, V., Chandrasekaran, A., Das, A., Lee, S., Batra, D., and Parikh, D., "Evaluating visual conversational agents via cooperative human-ai games," *CoRR*, vol. abs/1708.05122, 2017.

[42] Chattopadhyay, P., Yadav, D., Prabhu, V., Chandrasekaran, A., Das, A., Lee, S., Batra, D., and Parikh, D., "Evaluating visual conversational agents via cooperative human-ai games," in *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2017.

[43] Chen, D. and Manning, C. D., "A fast and accurate dependency parser using neural networks," in *EMNLP*, 2014.

[44] Chen, H., Branavan, S., Barzilay, R., Karger, D. R., and others, "Content modeling using latent permutations," *Journal of Artificial Intelligence Research*, 2009.

[45] Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., and Barnes, M., "Situation awareness-based agent transparency," tech. rep., ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD HUMAN RESEARCH AND ENGINEERING DIRECTORATE, 2014.

[46] Cho, K., Courville, A., and Bengio, Y., "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.

[47] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y., "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *EMNLP*, 2014.

[48] Choi, J., Oh, T.-H., and So Kweon, I., "Video-story composition via plot analysis," in *CVPR*, 2016.

[49] Chopra, S., Hadsell, R., and Lecun, Y., "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*, 2005.

[50] Clark, H. H., "Using language," 1996.

[51] Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., and Riedl, J., "Is seeing believing?: how recommender system interfaces affect users' opinions," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 585–592, ACM, 2003.

[52] Das, A., Agrawal, H., Zitnick, C. L., Parikh, D., and Batra, D., "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[53] Das, A., Agrawal, H., Zitnick, L., Parikh, D., and Batra, D., "Human attention in visual question answering: Do humans and deep networks look at the same regions?," *Computer Vision and Image Understanding*, vol. 163, pp. 90–100, 2017.

[54] Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D., "Visual Dialog," in *CVPR*, 2017.

[55] Davidov, D., Tsur, O., and Rappoport, A., "Semi-supervised recognition of sarcastic sentences in twitter and amazon," in *Conference on Computational Natural Language Learning*, 2010.

[56] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255, IEEE, 2009.

[57] der Maaten, L. V. and Hinton, G., "Visualizing data using t-SNE," *Journal of Machine Learning Research*, 2008.

[58] Deza, A. and Parikh, D., "Understanding image virality," in *CVPR*, 2015.

[59] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T., "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.

[60] Dragan, A. D., Lee, K. C., and Srinivasa, S. S., "Legibility and predictability of robot motion," in *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, pp. 301–308, IEEE, 2013.

[61] Dubey, R., Peterson, J., Khosla, A., Yang, M.-H., and Ghanem, B., "What makes an object memorable?," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1089–1097, 2015.

[62] Duran, R. L., "Communicative adaptability: A measure of social communicative competence," *Communication Quarterly*, 1983.

[63] Duran, R. L., "Communicative adaptability: A review of conceptualization and measurement.," *Communication Quarterly*, 1992.

[64] El Kaliouby, R. and Robinson, P., "Mind reading machines: Automated inference of cognitive mental states from video," in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 1, pp. 682–688, IEEE, 2004.

[65] El Kaliouby, R. and Robinson, P., "Real-time inference of complex mental states from facial expressions and head gestures," in *Real-time vision for human-computer interaction*, pp. 181–200, Springer, 2005.

[66] Engel, D., Woolley, A. W., Jing, L. X., Chabris, C. F., and Malone, T. W., "Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence equally well online and face-to-face," *PloS one*, 2014.

[67] Eysenbach, B., Vondrick, C., and Torralba, A., "Who is mistaken?," *arXiv preprint arXiv:1612.01175*, 2016.

[68] Fader, A., Zettlemoyer, L., and Etzioni, O., "Open question answering over curated and extracted knowledge bases," in *ACM SIGKDD*, 2014.

[69] Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., and others, "Visual storytelling," *arXiv preprint arXiv:1604.03968*, 2016.

[70] Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., and Mueller, E. T., "Watson: beyond jeopardy!," *Artificial Intelligence*, vol. 199, pp. 93–105, 2013.

[71] Ferrucci, D. A., "Introduction to this is watson," *IBM Journal of Research and Development*, vol. 56, no. 3.4, pp. 1–1, 2012.

[72] Firth, J. R., *A synopsis of linguistic theory*. Blackwell, 1957.

[73] Fouhey, D. F. and Zitnick, C. L., "Predicting object dynamics in scenes," in *CVPR*, 2014.

[74] Freud, S., *The Joke and Its Relation to the Unconscious*. Penguin, 2003.

[75] Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M., "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[76] Fussell, S. R., Kiesler, S., Setlock, L. D., and Yew, V., "How people anthropomorphize robots," in *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, pp. 145–152, IEEE, 2008.

[77] Fussell, S. R., Kiesler, S., Setlock, L. D., and Yew, V., "How people anthropomorphize robots," in *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, pp. 145–152, IEEE, 2008.

[78] Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., and Xu, W., "Are you talking to a machine? dataset and methods for multilingual image question," in *Advances in Neural Information Processing Systems*, pp. 2296–2304, 2015.

[79] Geman, D., Geman, S., Hallonquist, N., and Younes, L., "Visual turing test for computer vision systems," *Proceedings of the National Academy of Sciences*, vol. 112, no. 12, pp. 3618–3623, 2015.

[80] Ghazvininejad, M., Shi, X., Choi, Y., and Knight, K., "Generating topical poetry," in *EMNLP*, 2016.

[81] Goodchilds, J. D., Goldstein, J., and McGhee, P., "On being titty: Causes, correlates, and consequences," *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, 1972.

[82] Goodfellow, I. J., Shlens, J., and Szegedy, C., "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[83] Goyal, Y., Agrawal, A., Batra, D., and Parikh, D., "VQA Challenge Leaderboard." `http://www.visualqa.org/roe.html`, 2017.

[84] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D., "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," *arXiv preprint arXiv:1612.00837*, 2016.

[85] Goyal, Y., Mohapatra, A., Parikh, D., and Batra, D., "Towards transparent ai systems: Interpreting visual question answering models," *arXiv preprint arXiv:1608.08974*, 2016.

[86] Grosz, B., "What question would turing pose today?," *AI Magazine*, 2012.

[87] Gunning, D., "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[88] Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., and Van Gool, L., "The interestingness of images," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1633–1640, 2013.

[89] Hang, L., "A short introduction to learning to rank," *IEICE TRANSACTIONS on Information and Systems*, 2011.

[90] Harris, Z. S., "Distributional structure. word, 10 (2-3): 146–162. reprinted in fodor, j. a and katz, jj (eds.), readings in the philosophy of language," 1954.

[91] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[92] Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T., "Generating visual explanations," in *European Conference on Computer Vision*, pp. 3–19, Springer, 2016.

[93] Hitzeman, J., Moens, M., and Grover, C., "Algorithms for analysing the temporal structure of discourse," in *EACL*, 1995.

[94] Hochreiter, S. and Schmidhuber, J., "Long short-term memory," *Neural computation*, 1997.

[95] Horsky, J., Schiff, G. D., Johnston, D., Mercincavage, L., Bell, D., and Middleton, B., "Interface design principles for usable decision support: a targeted review of best practices for clinical prescribing interventions," *Journal of biomedical informatics*, vol. 45, no. 6, pp. 1202–1216, 2012.

[96] Hu, R., Dollár, P., He, K., Darrell, T., and Girshick, R., "Learning to segment every thing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4233–4241, 2018.

[97] Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Mishra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., and Mitchell, M., "Visual storytelling," in *NAACL*, 2016.

[98] Hurley, M. M., Dennett, D. C., and Adams, R. B., *Inside jokes: Using humor to reverse-engineer the mind*. MIT Press, 2011.

[99] Hutson, M., *The 7 laws of magical thinking: How irrational beliefs keep us happy, healthy, and sane*. Penguin, 2012.

[100] Hwang, C. H. and Schubert, L. K., "Tense trees as the fine structure of discourse," in *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 1992.

[101] Isola, P., Parikh, D., Torralba, A., and Oliva, A., "Understanding the intrinsic memorability of images," in *NIPS*, 2011.

[102] Jaech, A., Koncel-Kedziorski, R., and Ostendorf, M., "Phonological pun-derstanding," in *HLT-NAACL*, 2016.

[103] Joseph, J. and Friedman, L., "IBMs Watson Helps Employees Tackle Cancer ." `https://bestdoctors.com/blog/2017/01/10/press-release/`, 2017. [Online; accessed 17-March-2008].

[104] Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y., "Exploring the limits of language modeling," *arXiv preprint arXiv:1602.02410*, 2016.

[105] Jyothi, P. and Livescu, K., "Revisiting word neighborhoods for speech recognition," *ACL 2014*, 2014.

[106] Kamar, E., Gal, Y., and Grosz, B. J., "Incorporating helpful behavior into collaborative planning," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pp. 875–882, International Foundation for Autonomous Agents and Multiagent Systems, 2009.

[107] Kamp, H. and Reyle, U., "From discourse to logic; introduction to the model theoretic semantics of natural language," 1993.

[108] Kao, J. T., Levy, R., and Goodman, N. D., "The funny thing about incongruity: A computational model of humor in puns," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2013.

[109] Karpathy, A., "Neuraltalk2." `https://github.com/karpathy/neuraltalk2`, 2016. [Online; accessed 04-May-2017].

[110] Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. L., "Referitgame: Referring to objects in photographs of natural scenes.," in *EMNLP*, pp. 787–798, 2014.

[111] Kehler, A., "Coherence and the resolution of ellipsis," *Linguistics and Philosophy*, 2000.

[112] Khosla, A., Das Sarma, A., and Hamid, R., "What makes an image popular?," in *International Conference on World Wide Web*, 2014.

[113] Kim, G., Moon, S., and Sigal, L., "Joint photo stream and blog post summarization and exploration," in *CVPR*, 2015.

[114] Kim, G., Sigal, L., and Xing, E., "Joint summarization of large-scale collections of web images and videos for storyline reconstruction," in *CVPR*, 2014.

[115] Kim, G. and Xing, E., "Reconstructing storyline graphs for image recommendation from web community photos," in *CVPR*, 2014.

[116] Kim, J.-H., On, K.-W., Kim, J., Ha, J.-W., and Zhang, B.-T., "Hadamard product for low-rank bilinear pooling," *arXiv preprint arXiv:1610.04325*, 2016.

[117] Kiros, R., Salakhutdinov, R., and Zemel, R. S., "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[118] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S., "Skip-thought vectors," in *NIPS*, 2015.

[119] Kottur, S., Vedantam, R., Moura, J. M., and Parikh, D., "Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4985–4994, 2016.

[120] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.

[121] Krupenye, C., Kano, F., Hirata, S., Call, J., and Tomasello, M., "Great apes anticipate that other individuals will act according to false beliefs," *Science*, vol. 354, no. 6308, pp. 110–114, 2016.

[122] Kulesza, T., Stumpf, S., Burnett, M., and Kwan, I., "Tell me more?: the effects of mental model soundness on personalizing an intelligent agent," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1–10, ACM, 2012.

[123] Lapata, M. and Lascarides, A., "Learning sentence-internal temporal relations.," *Journal of Artificial Intelligence Research*, 2006.

[124] Lascarides, A. and Asher, N., "Temporal interpretation, discourse relations and commonsense entailment," *Linguistics and philosophy*, 1993.

[125] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L., "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, pp. 740–755, Springer, 2014.

[126] Lin, X. and Parikh, D., "Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks," in *CVPR*, 2015.

[127] Little, G., "How many turkers are there.(dec 2009)," 2009.

[128] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C., "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, pp. 21–37, Springer, 2016.

[129] Loper, E. and Bird, S., "Nltk: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, Association for Computational Linguistics, 2002.

[130] Lu, J., Lin, X., Batra, D., and Parikh, D., "Deeper lstm and normalized cnn visual question answering model." `https://github.com/VT-vision-lab/VQA_LSTM_CNN`, 2015.

[131] Lu, J., Yang, J., Batra, D., and Parikh, D., "Hierarchical question-image co-attention for visual question answering," in *Advances In Neural Information Processing Systems*, pp. 289–297, 2016.

[132] MacLeod, H., Bennett, C. L., Morris, M. R., and Cutrell, E., "Understanding blind peoples experiences with computer-generated captions of social media images."

[133] Magnini, B., Zanoli, R., Dagan, I., Eichler, K., Neumann, G., Noh, T.-G., Pado, S., Stern, A., and Levy, O., "The excitement open platform for textual inferences.," in *ACL (System Demonstrations)*, 2014.

[134] Mahapatra, A. and Srivastava, J., "Incongruity versus incongruity resolution," *2013 International Conference on Social Computing*, pp. 25–32, 2013.

[135] Malinowski, M. and Fritz, M., "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Advances in Neural Information Processing Systems*, pp. 1682–1690, 2014.

[136] Mani, I. and Schiffman, B., "Temporally anchoring and ordering events in news," *Time and Event Recognition in Natural Language. John Benjamins*, 2005.

[137] Mani, I., Schiffman, B., and Zhang, J., "Inferring temporal ordering of events in news," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, Association for Computational Linguistics, 2003.

[138] Mani, I., Verhagen, M., Wellner, B., Lee, C. M., and Pustejovsky, J., "Machine learning of temporal relations," in *COLING-ACL*, 2006.

[139] Martin, R. A. and Kuiper, N. A., "Daily occurrence of laughter: Relationships with age, gender, and type a personality," *Humor*, 1999.

[140] McGhee, P. E., "Chapter 5: The contribution of humor to children's social development," *Journal of Children in Contemporary Society*, 1989.

[141] McGraw, A. P. and Warren, C., "Benign violations making immoral behavior funny," *Psychological Science*, 2010.

[142] Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., and Procci, K., "Intelligent agent transparency in human–agent teaming for multi-uxv management," *Human factors*, vol. 58, no. 3, pp. 401–415, 2016.

[143] Meyer, R., "A New Caption That Works for Every New Yorker Cartoon." `https://www.theatlantic.com/notes/2015/09/a-new-universal-new-yorker-cartoon-caption-id-like-to-add-you-to-my-professi406783/`, 2015. [Online; accessed 14-April-2017].

[144] Microsoft, "Microsoft Cognitive Services Computer Vision API." https://www.microsoft.com/cognitive-services/en-us/computer-vision-api, 2017. [Online; accessed 17-March-2008].

[145] Mihalcea, R., "The multidisciplinary facets of research on humour," in *International Workshop on Fuzzy Logic and Applications*, 2007.

[146] Mihalcea, R. and Pulman, S., "Characterizing humour: An exploration of features in humorous texts," *Computational Linguistics and Intelligent Text Processing*, 2007.

[147] Mihalcea, R. and Strapparava, C., "Making computers laugh: Investigations in automatic humor recognition," in *EMNLP*, 2005.

[148] Mikolov, T., Chen, K., Corrado, G., and Dean, J., "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[149] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J., "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013.

[150] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J., "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[151] Miller, T., *Adjusting Sense Representations for Word Sense Disambiguation and Automatic Pun Interpretation*. PhD thesis, tuprints, 2016.

[152] Miller, T. and Gurevych, I., "Automatic disambiguation of english puns.," in *ACL (1)*, pp. 719–729, 2015.

[153] Minsky, M., *Society of mind*. Simon and Schuster, 1988.

[154] Mitchell, J. P., "Inferences about mental states," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 364, no. 1521, pp. 1309–1316, 2009.

[155] Mnih, V., Heess, N., Graves, A., and others, "Recurrent models of visual attention," in *Advances in neural information processing systems*, pp. 2204–2212, 2014.

[156] Mobbs, D., Greicius, M. D., Abdel-Azim, E., Menon, V., and Reiss, A. L., "Humor modulates the mesolimbic reward centers," *Neuron*, 2003.

[157] Modi, A., "Event embeddings for semantic script modeling," in *CoNLL*, 2016.

[158] Modi, A. and Titov, I., "Inducing neural models of script knowledge," in *CoNLL*, 2014.

[159] Moran, J. M., Rain, M., Page-Gould, E., and Mar, R. A., "Do i amuse you? asymmetric predictors for humor appreciation and humor production," *Journal of Research in Personality*, 2014.

[160] Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., and Allen, J., "A corpus and cloze evaluation for deeper understanding of commonsense stories," in *NAACL*, 2016.

[161] Mulder, M. P. and Nijholt, A., *Humor Research: State of the Art*. University of Twente, Centre for Telematics and Information Technology, 2002.

[162] Munkres, J., "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, 1957.

[163] Murstein, B. I. and Brust, R. G., "Humor and interpersonal attraction," *Journal of Personality Assessment*, 1985.

[164] Mutlu, B., Forlizzi, J., and Hodgins, J., "A storytelling robot: Modeling and evaluation of human-like gaze behavior," in *Humanoid robots, 2006 6th IEEE-RAS international conference on*, pp. 518–523, IEEE, 2006.

[165] Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., and Doshi-Velez, F., "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1802.00682*, 2018.

[166] Oliver, N. M., Rosario, B., and Pentland, A. P., "A bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[167] Padó, S., Noh, T.-G., Stern, A., Wang, R., and Zanoli, R., "Design and realization of a modular architecture for textual entailment," *Natural Language Engineering*, 2015.

[168] Parikh, D. and Grauman, K., "Implied feedback: Learning nuances of user behavior in image search," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 745–752, 2013.

[169] Park, D. H., Hendricks, L. A., Akata, Z., Schiele, B., Darrell, T., and Rohrbach, M., "Attentive explanations: Justifying decisions and pointing to the evidence," *arXiv preprint arXiv:1612.04757*, 2016.

[170] Passonneau, R. J., "A computational model of the semantics of tense and aspect," *Computational Linguistics*, 1988.

[171] Pelikan, H. R. and Broth, M., "Why that nao?: How humans adapt to a conventional humanoid robot in taking turns-at-talk," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4921–4932, ACM, 2016.

[172] Pepicello, W. J. and Green, T. A., *Language of riddles: new perspectives*. The Ohio State University Press, 1984.

[173] Petrovic, S. and Matthews, D., "Unsupervised joke generation from big data.," in *ACL*, 2013.

[174] Pickup, L., Pan, Z., Wei, D., Shih, Y., Zhang, C., Zisserman, A., Scholkopf, B., and Freeman, W., "Seeing the arrow of time," in *CVPR*, 2014.

[175] Plato, Hamilton, E., and Cairns, H., *The Collected Dialogues of Plato, Including the Letters*. Pantheon Books, 1961.

[176] Plester, B., "Healthy humour: Using humour to cope at work," *New Zealand Journal of Social Sciences Online*, 2009.

[177] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H., "Manipulating and measuring model interpretability," in *NIPS 2017 Transparent and Interpretable Machine Learning in Safety Critical Environments Workshop*, 2017.

[178] Premack, D. and Woodruff, G., "Does the chimpanzee have a theory of mind?," *Behavioral and brain sciences*, vol. 1, no. 04, pp. 515–526, 1978.

[179] Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and others, "The timebank corpus," in *Corpus linguistics*, 2003.

[180] Radev, D., Stent, A., Tetreault, J., Pappu, A., Iliakopoulou, A., Chanfreau, A., de Juan, P., Vallmitjana, J., Jaimes, A., and Jha, R., "Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest," *arXiv preprint arXiv:1506.08126*, 2015.

[181] Ramanathan, V., Tang, K., Mori, G., and Fei-Fei, L., "Learning temporal embeddings for complex video analysis," in *CVPR*, 2015.

[182] Raskin, V., "Semantic mechanisms of humor," in *Annual Meeting of the Berkeley Linguistics Society*, vol. 5, pp. 325–335, 1979.

[183] Ray, A., Christie, G., Bansal, M., Batra, D., and Parikh, D., "Question relevance in vqa: Identifying non-visual and false-premise questions," *arXiv preprint arXiv:1606.06622*, 2016.

[184] Recasens, A., Khosla, A., Vondrick, C., and Torralba, A., "Where are they looking?," in *Advances in Neural Information Processing Systems*, pp. 199–207, 2015.

[185] Recasens, M., de Marneffe, M.-C., and Potts, C., "The life and death of discourse entities: Identifying singleton mentions," in *HLT-NAACL*, 2013.

[186] Reddit, "Reddit."

[187] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[188] Regneri, M., Koller, A., and Pinkal, M., "Learning script knowledge with web experiments," in *ACL*, 2010.

[189] Ren, M., Kiros, R., and Zemel, R., "Exploring models and data for image question answering," in *NIPS*, 2015.

[190] Ren, S., He, K., Girshick, R., and Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[191] Ribeiro, M. T., Singh, S., and Guestrin, C., "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, ACM, 2016.

[192] Richardson, M., Burges, C. J., and Renshaw, E., "Mctest: A challenge dataset for the open-domain machine comprehension of text.," in *EMNLP*, 2013.

[193] Rinck, P., "Magnetic resonance in medicine. the basic textbook of the european magnetic resonance forum. 8th edition; 2014.."

[194] Ritchie, G., "Current directions in computational humour," *Artificial Intelligence Review*, 2001.

[195] Ritchie, G., "Can computers create humor?," *AI Magazine*, 2009.

[196] Ross, M. D., Owren, M. J., and Zimmermann, E., "Reconstructing the evolution of laughter in great apes and humans," *Current Biology*, vol. 19, no. 13, pp. 1106–1111, 2009.

[197] Ruch, W., Attardo, S., and Raskin, V., "Toward an empirical verification of the general theory of verbal humor.," *Humor: International Journal of Humor Research*, 1993.

[198] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, 2015.

[199] Salovey, P., Rothman, A. J., Detweiler, J. B., and Steward, W. T., "Emotional states and physical health.," *American Psychologist*, 2000.

[200] Savvy, T., "Watson will see you now: a supercomputer to help clinicians make informed treatment decisions," 2015.

[201] Scassellati, B., "Theory of mind for a humanoid robot," *Autonomous Robots*, vol. 12, no. 1, pp. 13–24, 2002.

[202] Schank, R. C. and Abelson, R. P., *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press, 2013.

[203] Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D., "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[204] Shahaf, D., Horvitz, E., and Mankoff, R., "Inside jokes: Identifying humorous cartoon captions," in *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[205] Shahaf, D., Horvitz, E., and Mankoff, R., "Inside jokes: Identifying humorous cartoon captions," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1065–1074, ACM, 2015.

[206] Sheppard, A., "Effect of mode of representation on visual humor," *Psychological Reports*, 1983.

[207] Sigurdsson, G. A., Chen, X., and Gupta, A., "Learning visual storylines with skipping recurrent neural networks," in *ECCV*, 2016.

[208] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., and others, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[209] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[210] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[211] Sinicropi, G., "La struttura della parodia— avvero: Bradamante in arli," *Strumenti Critici Torino*, 1981.

[212] Song, H., Nguyen, A.-D., Gong, M., and Lee, S., "A review of computer vision methods for purpose on computer-aided diagnosis," *Journal of International Society for Simulation Surgery*, vol. 3, no. 1, pp. 1–8, 2016.

[213] Spearman, C., "The proof and measurement of association between two things," *The American journal of psychology*, 1904.

[214] Stock, O. and Strapparava, C., "HAHAcronym: A computational humor system," in *ACL*, 2005.

[215] Suls, J. M., "A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis," *The Psychology of Humor: Theoretical Perspectives and Empirical Issues*, 1972.

[216] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A., "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.

[217] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[218] Tang, K., Fei-Fei, L., and Koller, D., "Learning latent temporal structure for complex event detection," in *CVPR*, 2012.

[219] Taylor, J. and Mazlack, L., "Computationally recognizing wordplay in jokes," *Proceedings of CogSci*, 2004.

[220] Trottier, D., *The screenwriter's bible: A complete guide to writing, formatting, and selling your script.* Silman-James Press, 1998.

[221] Tsakona, V., "Language and image interaction in cartoons: Towards a multimodal theory of humor," *Journal of Pragmatics*, 2009.

[222] Tu, K., Meng, M., Lee, M. W., Choe, T. E., and Zhu, S.-C., "Joint video and text parsing for understanding events and answering queries," *IEEE MultiMedia*, vol. 21, no. 2, pp. 42–70, 2014.

[223] Vedantam, R., Lin, X., Batra, T., Zitnick, C. L., and Parikh, D., "Learning common sense through visual abstraction," in *ICCV*, 2015.

[224] Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R., "Order-embeddings of images and language," in *ICLR*, 2016.

[225] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D., "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

[226] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D., "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[227] Vondrick, C., Oktay, D., Pirsiavash, H., and Torralba, A., "Predicting motivations of actions by leveraging text," in *Proceedings of the IEEE Conference on Comslputer Vision and Pattern Recognition*, pp. 2997–3005, 2016.

[228] Vondrick, C., Pirsiavash, H., and Torralba, A., "Anticipating visual representations from unlabeled video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 98–106, 2016.

[229] Wang, S. I., Liang, P., and Manning, C. D., "Learning language games through interaction," *arXiv preprint arXiv:1606.02447*, 2016.

[230] Wang, W. Y., Mehdad, Y., Radev, D. R., and Stent, A., "A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization," in *NAACL*, 2016.

[231] Wang, W. Y. and Wen, M., "I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions," in *NAACL*, 2015.

[232] Wanzer, M. B., Booth-Butterfield, M., and Booth-Butterfield, S., "Are funny people popular? an examination of humor orientation, loneliness, and social attraction," *Communication Quarterly*, 1996.

[233] Watson, K. K., Matthews, B. J., and Allman, J. M., "Brain activation during sight gags and language-dependent humor," *Cerebral Cortex*, 2007.

[234] Wayland, M., "Nissan self-driving system teams AI with human advisers." `http://www.detroitnews.com/story/business/autos/foreign/2017/01/05/nissan-sam/96224020/`, 2017. [Online; accessed 17-March-2008].

[235] Webber, B. L., "Tense as discourse anaphor," *Computational Linguistics*, 1988.

[236] Weston, J., Bordes, A., Chopra, S., and Mikolov, T., "Towards AI-complete question answering: A set of prerequisite toy tasks," *arXiv preprint arXiv:1502.05698*, 2015.

[237] Wikipedia, "Spearman's rank correlation." `https://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient`. Accessed: 05-19-2016.

[238] Wikipedia, "Theories of humor." `https://en.wikipedia.org/wiki/Theories_of_humor`. Accessed: 26-Nov-2018.

[239] Wikipedia, "Humor," November 2015.

[240] Wikipedia, "Hungarian algorithm." `https://en.wikipedia.org/wiki/Hungarian_algorithm`, Accessed: 06-03-2016.

[241] Wimmer, H. and Perner, J., "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception," *Cognition*, vol. 13, no. 1, pp. 103–128, 1983.

[242] Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., and Malone, T. W., "Evidence for a collective intelligence factor in the performance of human groups," *science*, 2010.

[243] Wu, J. and Mooney, R. J., "Faithful multimodal explanation for visual question answering," *arXiv preprint arXiv:1809.02805*, 2018.

[244] Wu, S., Wieland, J., Farivar, O., and Schiller, J., "Automatic alt-text: Computer-generated image descriptions for blind users on a social network service," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 1180–1192, ACM, 2017.

[245] Xu, H. and Saenko, K., "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *European Conference on Computer Vision*, pp. 451–466, Springer, 2016.

[246] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y., "Show, attend and tell: Neural image caption generation with visual attention.," in *ICML*, vol. 14, pp. 77–81, 2015.

[247] Yang, D., Lavie, A., Dyer, C., and Hovy, E. H., "Humor recognition and humor anchor extraction," in *EMNLP*, 2015.

[248] Yang, Z., He, X., Gao, J., Deng, L., and Smola, A., "Stacked attention networks for image question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 21–29, 2016.

[249] Yu, Z., Nicolich-Henkin, L., Black, A. W., and Rudnicky, A. I., "A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement," in *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 55, 2016.

[250] Zeiler, M. D. and Fergus, R., "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Springer, 2014.

[251] Zhang, P., Wang, J., Farhadi, A., Hebert, M., and Parikh, D., "Predicting failures of vision systems," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3566–3573, 2014.

[252] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., "Object detectors emerge in deep scene cnns," *arXiv preprint arXiv:1412.6856*, 2014.

[253] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

[254] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S., "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 19–27, 2015.

[255] Zitnick, C. L., Vedantam, R., and Parikh, D., "Adopting abstract images for semantic scene understanding," *PAMI*, 2014.

[256] Zitnick, C. L., Agrawal, A., Antol, S., Mitchell, M., Batra, D., and Parikh, D., "Measuring machine intelligence through visual question answering," *AI Magazine*, vol. 37, no. 1, pp. 63–72, 2016.

[257] Zitnick, C. L. and Parikh, D., "Bringing semantics into focus using visual abstraction," in *CVPR*, 2013.

[258] Zitnick, C. L., Parikh, D., and Vanderwende, L., "Learning the visual interpretation of sentences," in *ICCV*, 2013.

[259] Ziv, A. and Gadish, O., "Humor and marital satisfaction," *The Journal of Social Psychology*, 1989.

[260] Zwicky, A. and Zwicky, E., "Imperfect puns, markedness, and phonological similarity: With fronds like these, who needs anemones," *Folia Linguistica*, 1986.